IPWE for EMBARC

2020-04-08

Covariates in the EMBARC dataset

Covariate name	Description
w0_4165	A not B Interference Reaction Time in negative trials
w0_4167	A not B Interference Reaction Time in non-negative trials
w0_4163	A not B Interference Reaction Time in all trials
w0_4162	A not B Itotal number of correct trials
w0_4169	Median Reaction time for correct trials in the Choice reaction time task
w0_1844	Number of valid recalled words in the Word Fluency task
w0_1916	Flanker Accuracy, an Accuracy effect is a measure of interference effects; Higher scores are indicative of increased interference effects (i.e., reduced cognitive control).
w0_1915	Flanker Reaction Time, a measure of interference effects; Higher scores are indicative of increased interference effects (i.e., reduced cognitive control).
w0_1920	Accuracy effect, it measures post-conflict behavioral adjustments; Higher values indicate better cognitive control
age_evaluation	Age at baseline
sex	Sex
hamd17_baseline	Severity of depression at baseline
dur_MDE	Duration of current major depressive episode
age_MDE	Age of first major depressive episode
Greg_FH	Family history of psychotic and depressive disorder (4 ordered levels)
fatigue	level of fatigue at baseline
hypersomnia	presence/absence of hypersomnia at baseline
axis2	Severity of the most severy Axis II diagnosis at baseline
anger_attack	Severity of anger attacks at baseline
anxious	Severity of anxiety at baseline

Change score

For each subject in the data, we have their outcome score at week 0,1,2,3,4,6,8 week. The change score in the dataset is calculated as the difference between the score at the week 8 and the score at the week 0.

$$\text{change score: } Y(8) - Y(0)$$

(160 subjects in total, 87 in placebo group and 73 in drug group. All of them have scores at week 8).

IPWE

We estimate the overall outcome, e.g., change score for EMBARC data by using the inverse probablity weighted estimator, following the Zhang's paper.

In the paper, the IPWE is the estimator of $E(Y^*)$, defined as

$$IPWE(\eta) = \frac{1}{n}\sum_{i=1}^{n}\frac{C_{\eta,i}Y_i}{\pi(X_i;\eta,\gamma)}$$

where

- $C = Ag(X,\eta) + (1-A)(1-g(X,\eta)$
- A is the observed received treatment, $A = 0$ or 1;
- $g()$ function projecting covariates X to treatment assignment, $g(X) = 1$ or 0.
- η is the parameter; $\pi(X_i;\eta,\gamma)$
- $\pi(X_i;\eta,\gamma)$ is the propensity score. Since we look at randomized trials, it is a constant.

Therefore, $C =$

Subject i	True drug group	True placebo group
Estimated drug group	1	0
Estimated placebo group	0	1

That is, when $C = 1$, the counterfacutal outcome is observed; $C = 0$ otherwise.

He also defined AIPWE to avoid the misspecification of $\pi(X_i;\eta,\gamma)$, since our data is from a randomized trial, we do not need to use AIPWE.

However, in the formula, which n should we use to estimate the outcome mean?

$$\begin{cases} n = & \text{total number of subjects} \\ n = n^* = \sum_{i=1}^{n} C_i & \text{subjects whose counterfacutal outcomes are observed} \end{cases}$$

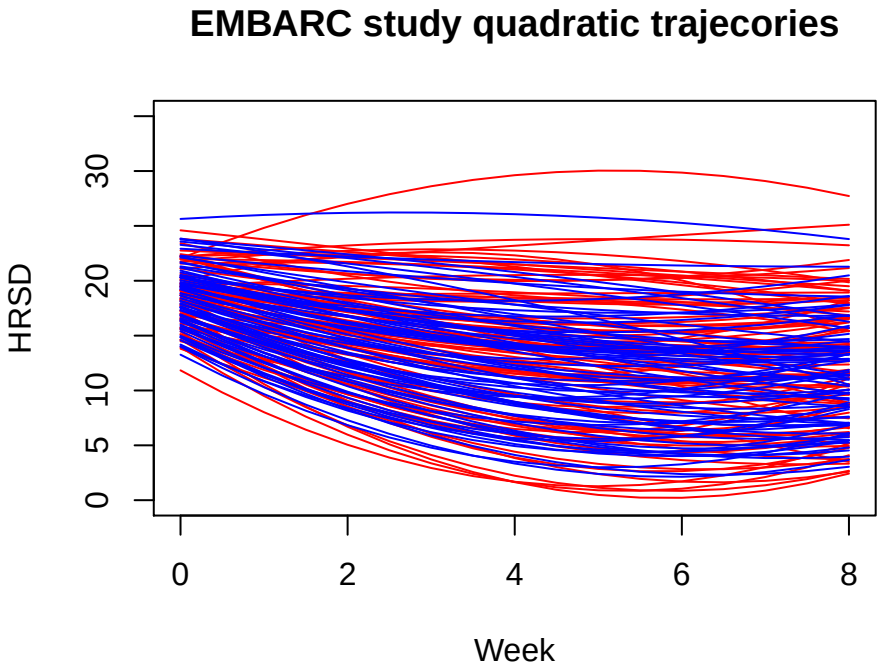
In our previous simulation study, since 1) we know the true counterfactual outcomes (drug and placebo group) and 2) we only looked at the direction of group assignement instead of the exact value of the change scores, we did not meet the problem proposed by Dr Frank Harrell.

However, when we analyze the EMBARC data, that problem pops up, i.e. we may need to consider the association between the change score and baseline score. In other words, a larger change score doesn not always imply a better treatment effect, since outcome of people with low baseline score will not change a lot.

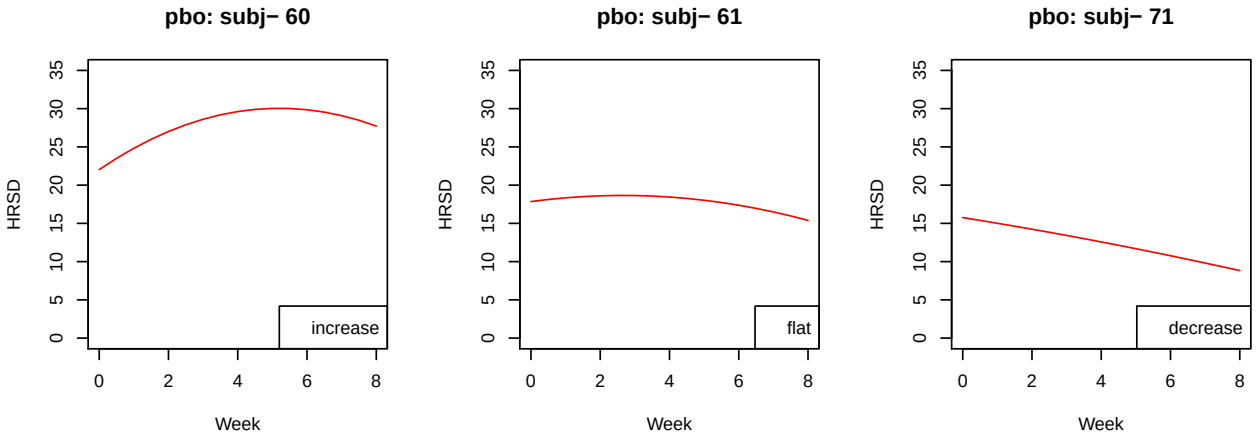
Also, another problem pops up, that is, the shapes of the trajectories are different, i.e. some are concave, some are convex.

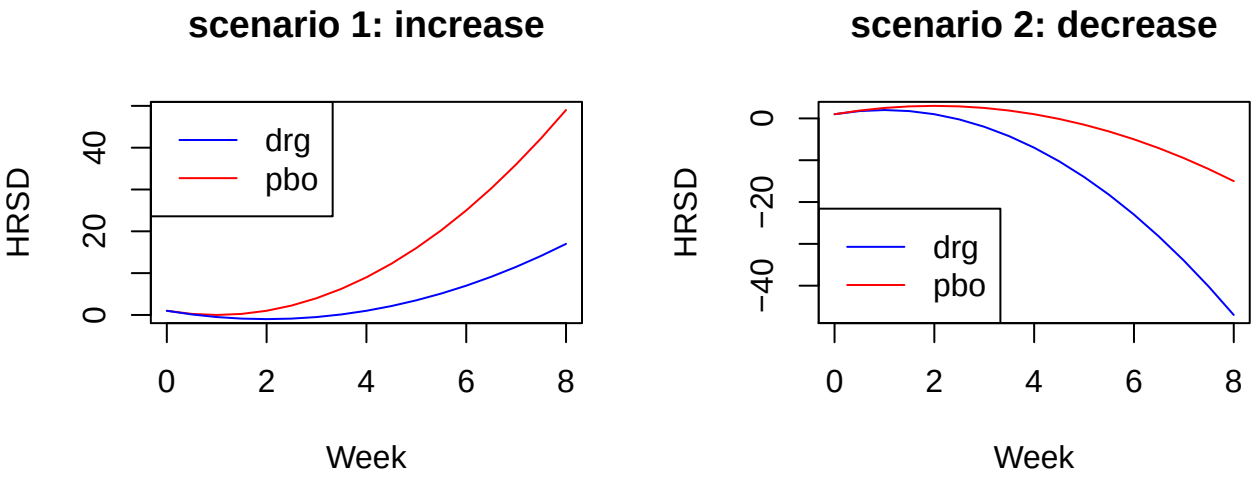
For example:

Baseline score affect



Trajectories affect





And therefore, since the trajectories are different, we may include some increased outcomes and then may bring problem. For example, in the following table,

id	csdrg	cspbo	truegroup	assign	changescore	group1	group2	cs1	cs2
1	40	45	2	2	40	2	1	40	0
2	-3	-2	2	1	-2	2	1	0	-2
3	-10	-17	1	2	-10	1	2	0	-10
4	-5	-7	1	1	-7	1	2	-7	0

Calculate the IPWE for EMBARC data

To predict the group assignment on the EMBARC, we conduct a 10 fold cross validation. The EMBARC dataset is separated into 10 parts (each part have the same proprotion of treatment, e.g. 50% drug, 50% placebo). Each time we use 9 out of 10 parts as the training set and then rest 1 part as the testing set.

To train the data, we conduct:

- the change score method (linear regression)
- the longitudinal method applied with Kullback Leibler divergence
- all the baseline covariates in the EMBARC are included.

For the test dataset, after training the data, we can predict the outcomes for subjects in the test dataset, if they are taking drug or if they are taking placebo ($\hat{Y}_{drg}, \hat{Y}_{pbo}$). The associated estimated change score can be calculated. Then we may evaluate which subject should be assigned into drug group or which one should be in placebo group.

The we can calculate the IPWE

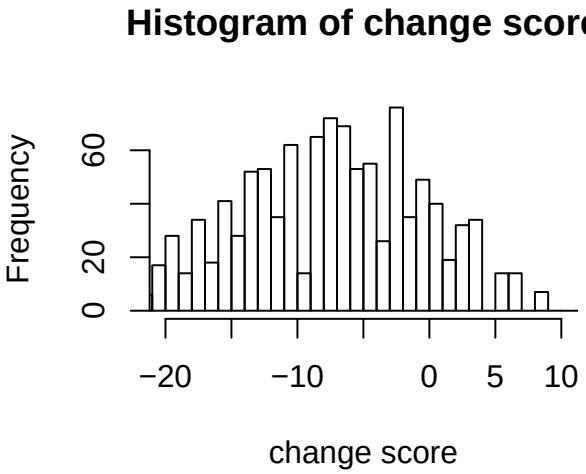
$$IPWE = \frac{1}{n}\sum_{i=1}^{n} C_{\eta,i} Y_i$$

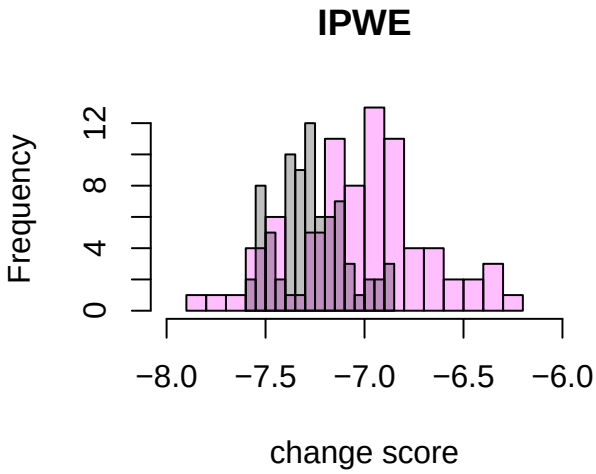
where $C_i = 1$ if the estimated assigned group = observed assigned group; $C_i = 0$ otherwise. Y_i is the observed change score in the test dataset.

After 10 times of CV, we get 10 IPWEs. And we then calculte the mean value of IPWE over the 10 fold CV.

The above procedures are repeated for 100 times.

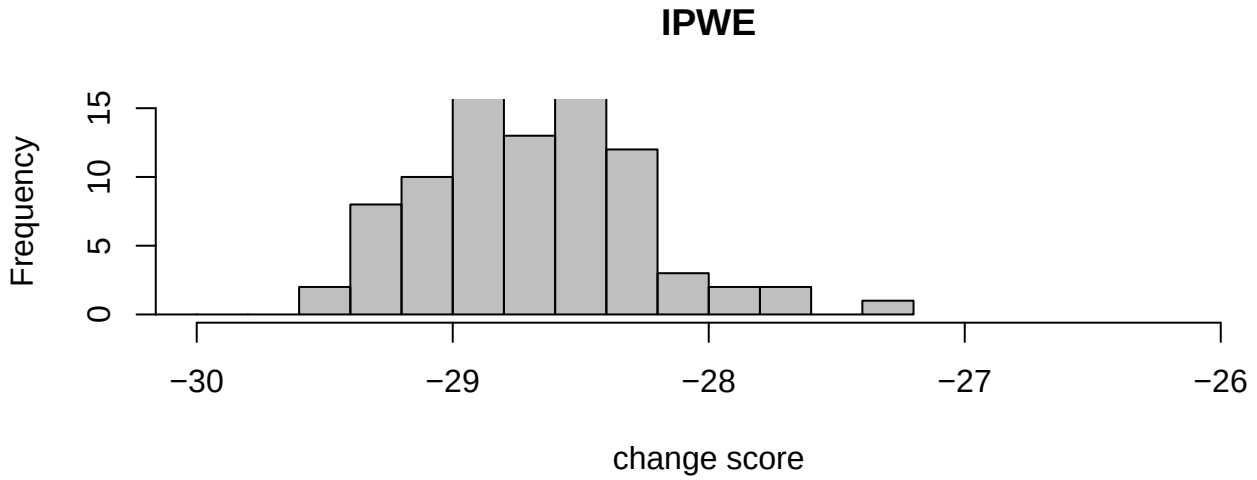
Results





	change score	KL-LME	linear
mean	-6.749	-7.273	-7.028
sd	7.018	0.173	0.339

Simulation



	KL-LME	linear

mean	-31.547	-28.702
sd	0.368	0.436