Environmental Moderators of Genetic Risk for Depression among Older US Adults: Evidence from the Wisconsin Longitudinal Study

Jinyuan Qi

OPR, Princeton University

Oct 4th

What do we know?

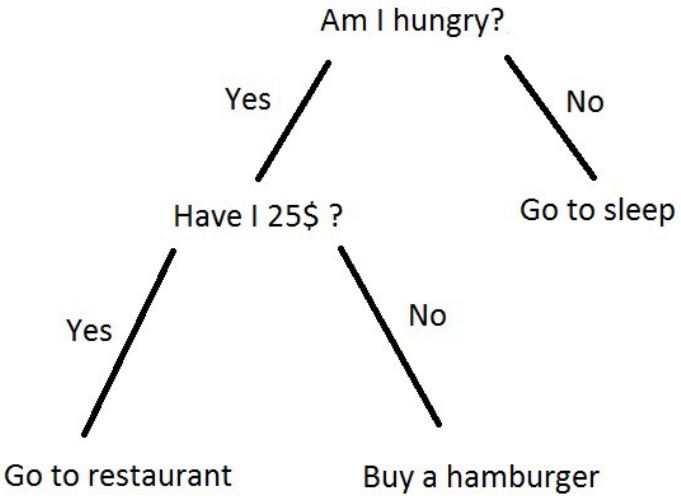
- Increasing burden of depression among the older adults in the US.
- Current understanding of depression? Gene-environment interaction?
- Interaction terms in linear regression - products of original variables
- Two-way interaction: $\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
- Three-way interaction:
  $\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3 + \beta_7 X_1 X_2 X_3$

What do we not know?

- Failing to account for interaction effects
- Versus introducing too many interaction terms
- Large $\boldsymbol{p}$, Small $\boldsymbol{n}$ problems in longitudinal survey data (e.g., WLS has 10,000+ graduates and 15,000+ variables).
- Exploratory data analysis - machine learning methods can help select important predictors and identify important interactions.

Decision trees - Model-based recursive partitioning

- Tree: leaves represent class labels & branches represent conjunctions of features that lead to those class labels.



- Split function chooses the best feature & the best value for that feature to minimize cost function (e.g., MSE)
- Recursive partitioning function

Random Forest

- Tree models are unstable
- Predictions are very sensitive to small changes to inputs.
- Random forests for variable selection: multiple trees constructed on random samples achieved either through bootstrapping or sub-sampling.

The `mobForest` package

The `mobForest` package constructs large number of model-based trees and the predictions are aggregated across these trees resulting in more stable predictions with variable importance.

Data - Wisconsin Longitudinal Study

- A one-third sample of all 1957 Wisconsin high school graduates (Three waves - 1993, 2004, & 2011)
- **Outcome:** a summary CES-D score (0-60) for depression that uses consistent CES-D questions.
- **PGS**: MTAG (Multi-Trait Analysis of GWAS)-based Polygenic Scores for Depression were created by Turley et al (2018)[1] using LDpred.

$$\bar{y} = \sum_{j=1}^{k} x_{ij} w j$$

- **Various risk factors:** e.g., education, marital status, wealth, etc.

[1] Turley et al. "Multi-trait analysis of genome-wide association summary statistics using MTAG". . In: *Nature Genetics* 50.2 (2018), pp. 229–237.

Regressors' coefficients in MOB Tree

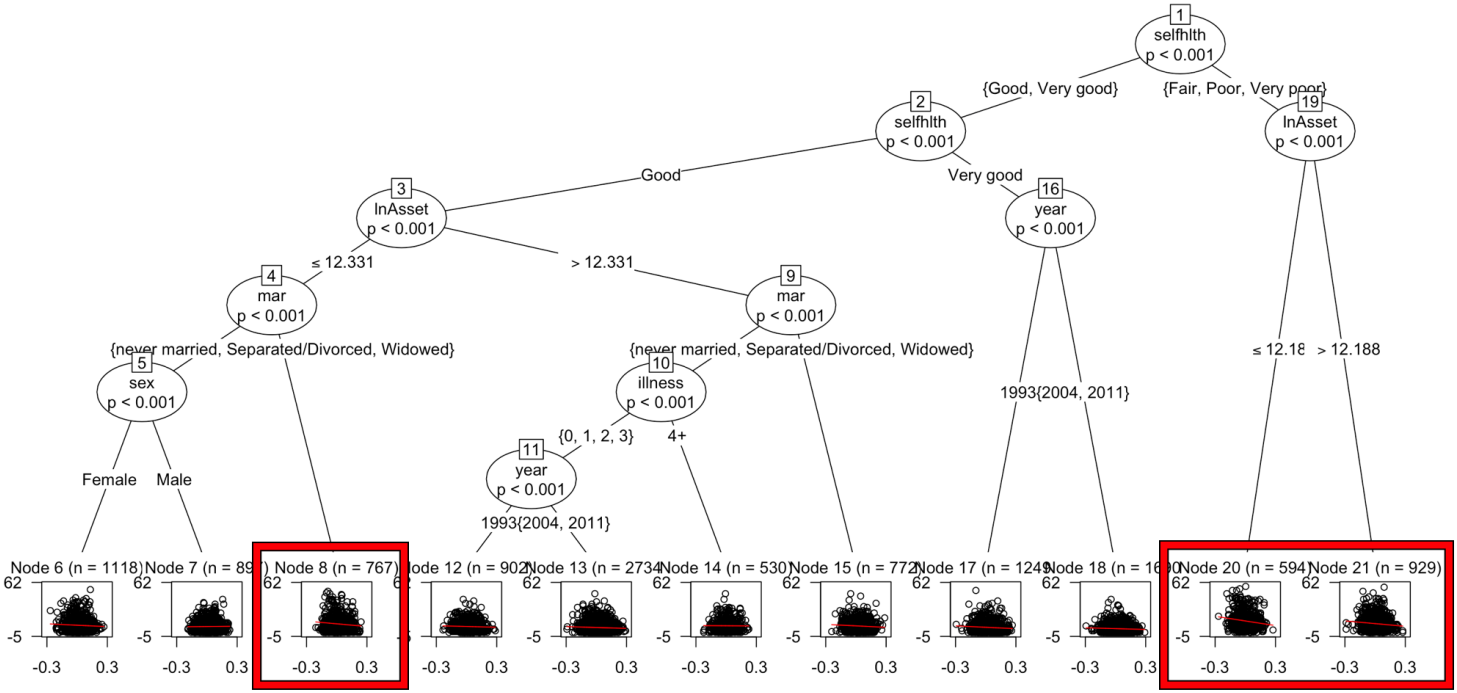
First MOB Tree - 45 nodes		
Node	(Intercept)	PGS
7	9.662***	-7.682
8	7.928***	-4.214
9	7.254***	1.159
10	10.632***	-11.164**
15	7.171***	-3.697
16	10.483***	-16.327
19	6.880***	18.635**
20	6.160***	-4.706**
22	6.376***	-7.709*
23	4.616***	-4.330
25	8.258***	-5.645
26	6.815***	-0.480
28	8.399***	25.692***
29	8.578***	-11.494**
33	5.994***	-4.412*
34	10.858***	-22.463
35	10.990***	-31.153*
37	4.408***	-3.283*
38	6.413***	-4.553
40	14.174***	-19.225
42	12.122***	-10.694
44	9.366***	-10.522**
45	14.241***	-18.690

Pruned MOB Tree - 21 nodes		
Node	(Intercept)	PGS
6	8.872***	-5.844*
7	7.254***	1.159
8	10.632***	-11.164**
12	7.322***	-3.079
13	6.067***	-3.664**
14	8.203***	-0.804
15	8.087***	-6.051*
17	6.332***	-5.824**
18	4.660***	-3.404*
20	14.174***	-19.225***
21	10.886***	-11.241***

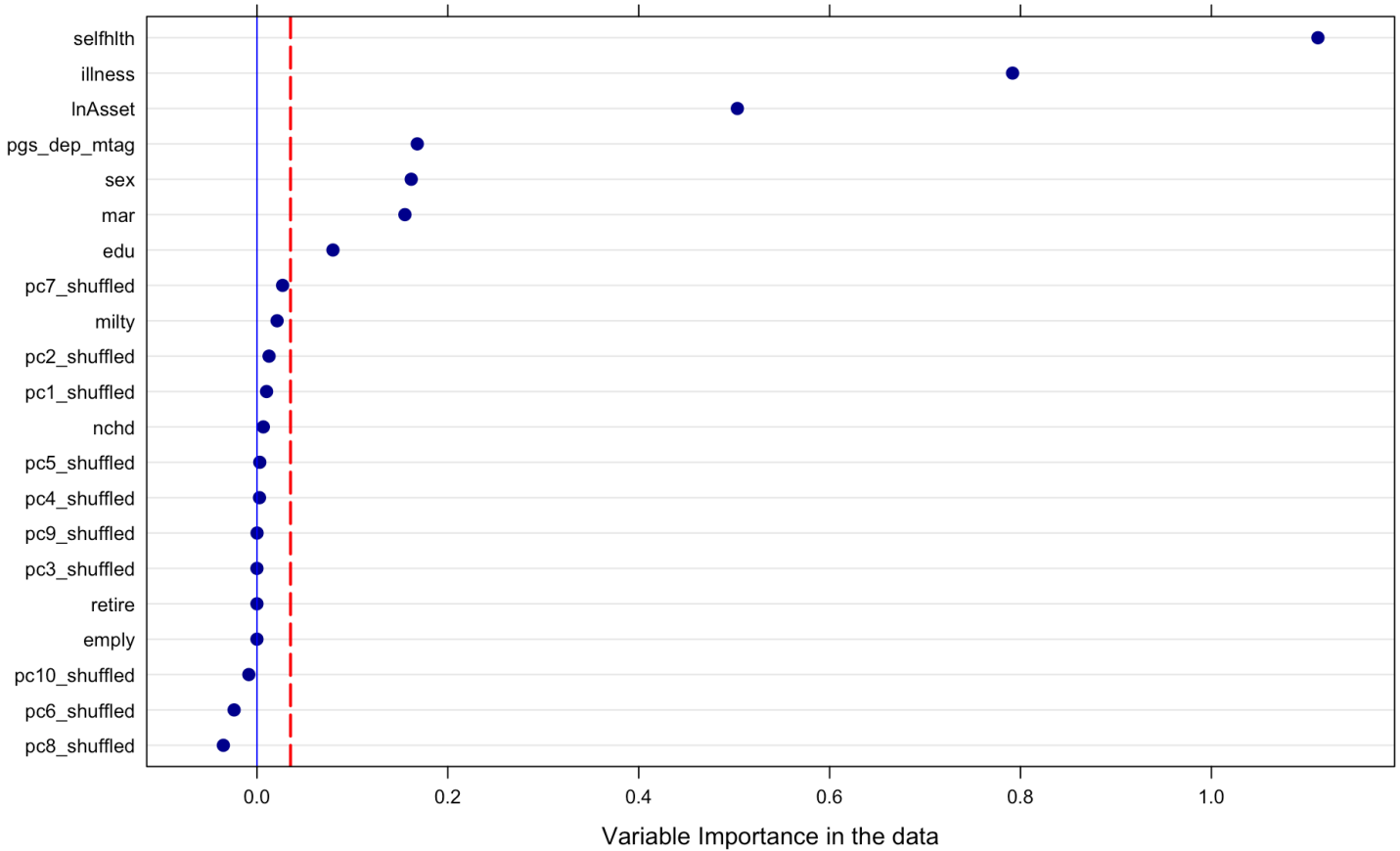
Pruned Tree including wave as regressor - 17 nodes		
Node	(Intercept)	PGS
7	9.864***	-7.152*
8	7.842***	-2.796
9	7.256***	-2.651*
10	10.857***	-1.388
11	11.173***	-8.474***
13	5.941***	-4.424***
14	8.121***	-4.384
16	15.125***	-14.330**
17	11.661***	-16.467***
Same Tree as above		
Node	Y2004	Y2011
7	-3.443***	-1.257
8	-1.426***	-1.568***
9	-1.637***	-0.781**
10	-2.929**	-1.938*
11	-2.515***	-2.311***
13	-1.667***	-1.393***
14	-2.217***	-2.422***
16	-2.949**	-1.208
17	-2.425**	0.322

Note: *** <0.001 ** <0.01 * <0.05

Tree example: Pruned MOB Tree - 21 nodes



Variable Selection from Random Forest



Discussion

- Self-rated health and physical illness and assets have greater importance in predicting CES-D scores than other variables for selection.
- The negative association is especially significant among those more disadvantaged subgroups with worse self-rated health, more illnesses, unmarried and less assets with higher intercept values.
- Gene-environment interactions could be very complex and multidimensional with multiple pathways and offsetting effects.
- Simply association - no causal inference (e.g., confounder, collider, etc.)
- The machine learning methods can be combined with other statistical methods to discover patterns and select features.

Thank you!

Model-Based Recursive Partitioning

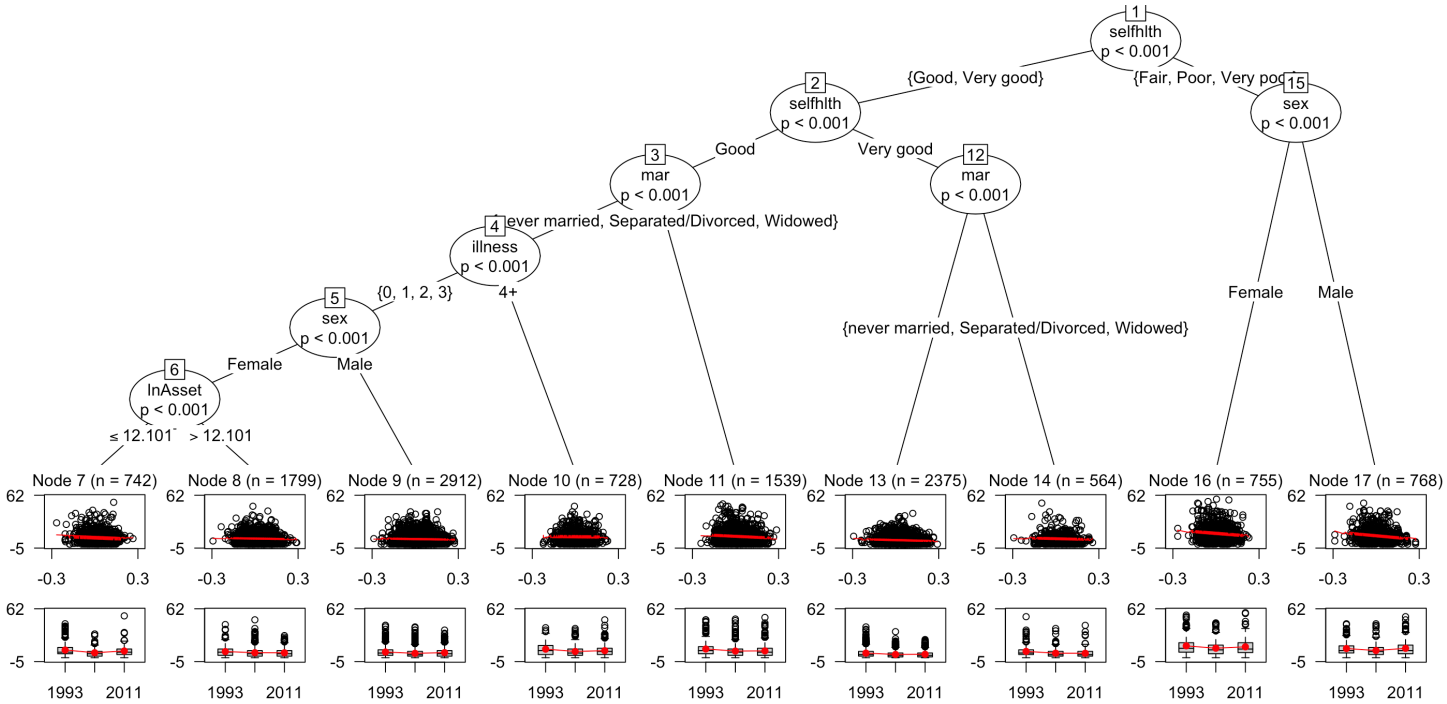
Grow a tree to explore different associations between PGS and depression among different subgroups:

1. Fit a parametric model to all
2. Test for parameter instability
3. If there is some overall parameter instability, split the model by the variable with the highest instability
4. Repeat the procedure at each of the child nodes.

The `party` package with `mob`

- Check overall instability - the minimal p value $< \alpha$ level
- Greedy algorithm - models are fit at every conceivable split point exhaustively

Tree example: Pruned MOB Tree including *wave* as regressor - 17 nodes



Decision trees[2] - recursive partitioning algorithm

- Split function chooses the best feature & the best value for that feature:

$$(j^*, t^*) = \arg \min_{j \in \{1,\ldots,D\}} \min_{t \in T_j} cost(\{x_i, y_i : x_{ij} \leq t\}) + cost(\{x_i, y_i : x_{ij} > t\})$$

- Pseudo codes for recursive partitioning function:

```
1  function fitTree(node, 𝒟, depth) ;
2  node.prediction = mean(y_i : i ∈ 𝒟) // or class label distribution ;
3  (j*, t*, 𝒟_L, 𝒟_R) = split(𝒟);
4  if not worthSplitting(depth, cost, 𝒟_L, 𝒟_R) then
5      return node
6  else
7      node.test = λx.x_{j*} < t* // anonymous function;
8      node.left = fitTree(node, 𝒟_L, depth+1);
9      node.right = fitTree(node, 𝒟_R, depth+1);
10     return node;
```

[2] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.