Submitted to the Annals of Applied Statistics

SUPPLEMENT TO "FEATURE SELECTION FOR GENERALIZED VARYING COEFFICIENT MIXED-EFFECT MODELS WITH APPLICATION TO OBESITY GWAS": CAUSAL INFERENCE

By Wanghuan Chu Runze Li Jingyuan Liu and Matthew Reimherr

While most GWAS studies focus on association analysis between genetic markers and phenotypes, the question of real interest is: does this genetic marker cause or increase the risk of a certain disease? This question falls naturally into the field of causal inference. The golden rule of making causal inference is using randomized experiment, but in many situations data can only be obtained from observational studies. The genotypes of SNPs, AA, Aa and aa, for instance, cannot be randomly assigned to any person. Over the past several decades, there have been substantial developments on how to infer causal effects from observational data by using statistical techniques to adjust for confounding factors. In this section, we apply two popular causal inference techniques, propensity score modeling and inverse probability weighting, to assess the causal effects of SNPs on BMI.

Propensity scores were first proposed by Rosenbaum and Rubin (1983), which aims to minimize the differences in confounding factors between the treatment and control groups. The score is defined as the conditional probability of being assigned to the treatment group given a set of observed potential confounders. Conditioning on the propensity score, the binary treatment assignment and the observed covariates are independent. In practice, propensity scores for binary treatments are usually estimated by logistic regression (Rosenbaum and Rubin (1984, 1985)). Imai and Van Dyk (2004) generalized the use of propensity scores to ordinal and categorical treatments, which can be applied to our data.

For each selected SNP, SNP_j, with three genotypes AA, Aa and aa, we estimate the propensity scores $e_g(\text{SNP}_j) = prob(\text{SNP}_j = g|\mathbf{X}^{(-j)})$, where $g \in \{AA, Aa, aa\}$ and $\mathbf{X}^{(-j)} = \{\text{SNP}_k : k \neq j, k \in \widehat{M}^{(f)}_{\tau_n}\}$. Then the inverse probability weighting technique is applied to the following model:

$$\begin{aligned}\text{BMI}^*_{ij} = \beta_0(\text{age}_{ij}) + \beta_1(\text{age}_{ij})\text{Gender}_i + \beta_2(\text{age}_{ij})\text{Smoke}_{ij} + \beta_3(\text{age}_{ij})\text{Alcohol}^*_{ij} \\ + \beta_4(\text{age}_{ij})\text{Alcohol}^{*2}_{ij} + \gamma_{k1}(\text{age}_{ij})I^{Aa}_{ik} + \gamma_{k2}(\text{age}_{ij})I^{aa}_{ik} + \varepsilon_{ij},\end{aligned}$$

where I^{Aa}_{ik} and I^{aa}_{ik} are the indicator functions of Aa and aa. Specifically, the weighted least squares estimates are computed for each selected SNP_j in $\widehat{M}_{\tau_n}^{(f)}$, with the weights $p(g_{ij})/e_{g_{ij}}(\mathrm{SNP}_{ij})$ assigned to the ith subject, where SNP_{ij} is the jth SNP for the ith subject, g_{ij} is the genotype of SNP_{ij}, and $p(g_{ij})$ is the proportion of g_{ij} in the population. Figure 1 shows the top 15 SNPs with smallest weighted residual sums of squares (WRSS).

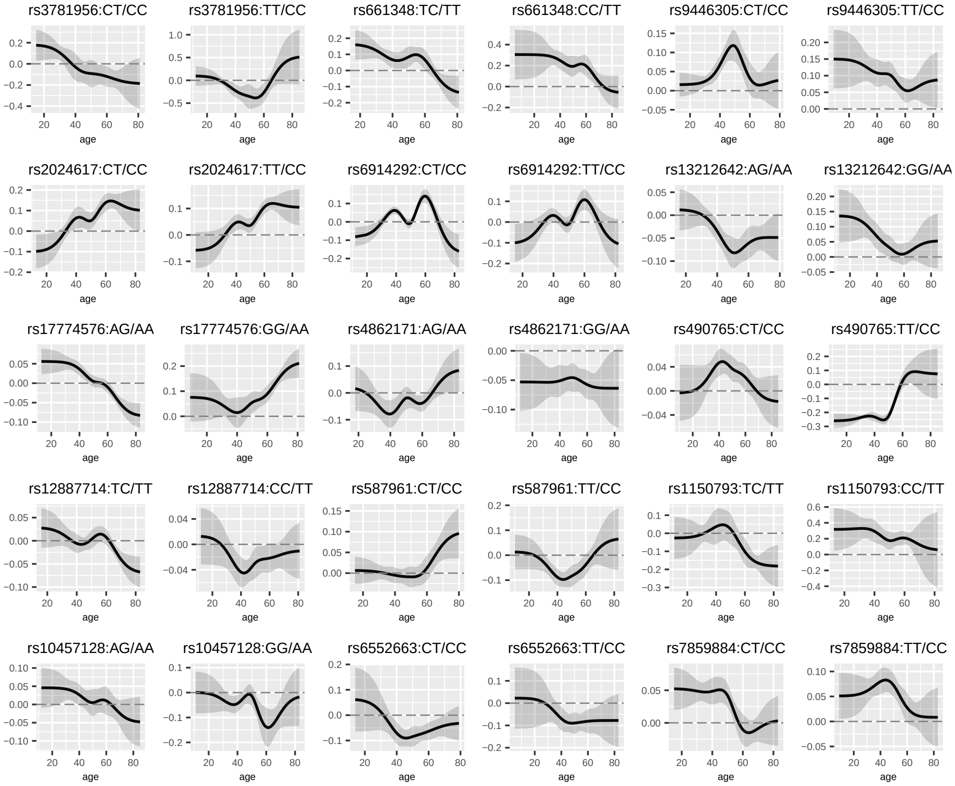

Fig 1: Causal effects of the top 15 SNPs from our empirical analysis

Furthermore, we conduct the same causal inferences for the 198 SNPs collected from previous research in literature[*]. The time-varying effects of the top 15 SNPs, with detailed information shown in Table 1, are depicted in Figure 2. Note that 11 of them are located on the well known "fat gene" FTO[†]. In Figure 3, we compare the top 2 SNPs in our study (the first row) and those in literature (the second row). We observe that the two SNPs obtained from our procedures have much larger causal effects.

[*] http://snpedia.com

[†] https://en.wikipedia.org/wiki/FTO_gene

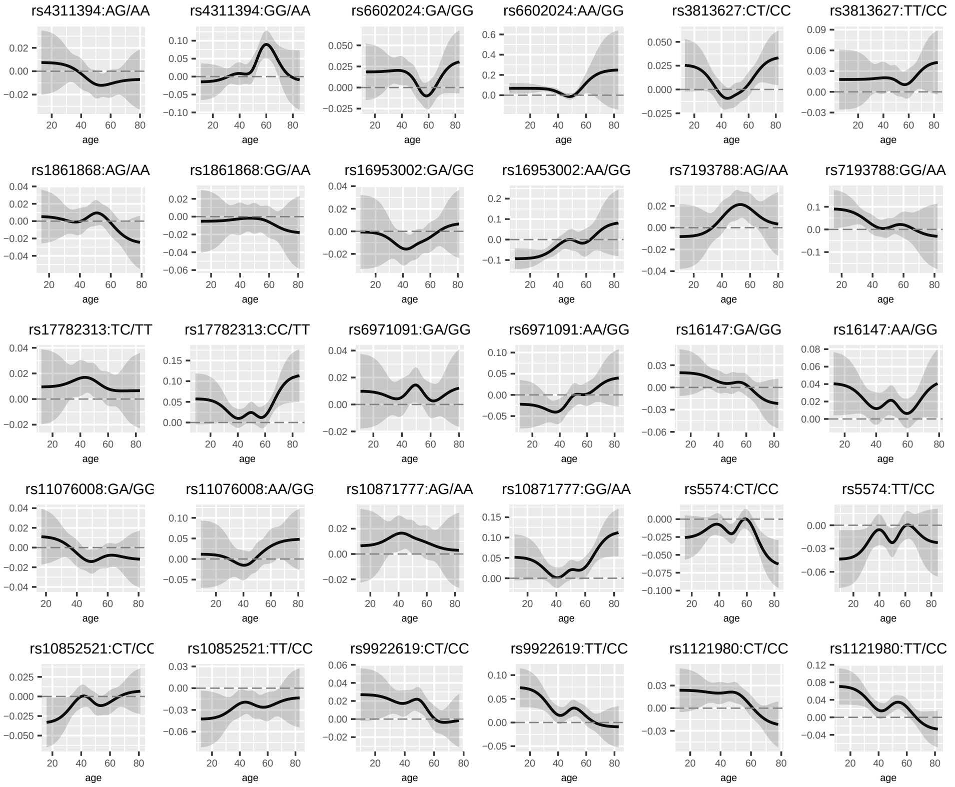

Fig 2: Causal effects of the top 15 SNPs from published research

Table 1
15 SNPs from previous research for BMI and obesity

SNP	Chromosome	Position	Gene	Risk allele	MAF
rs4311394	5	54004832	ARL15	G	25.57%
rs6602024	5	54004832	ARL15	A	9.33%
rs3813627	1	161225358	APOA2	T	34.64%
rs1861868	16	53756490	FTO	G	49.84%
rs16953002	16	54080912	FTO	A	16.29%
rs7193144	16	53776774	FTO	G	14.81%
rs17782313	18	60183864	MC4R	C	21.73%
rs6971091	7	128723233	FAM71F1	A	22.53%
rs16147	7	24283791	NPY	A	49.30%
rs11076008	16	53893411	FTO	A	21.18%
rs10871777	18	60184530	MC4R	G	22.04%
rs5574	7	24289514	NPY	T	46.75%
rs10852521	16	53771053	FTO	T	47.66%
rs9922619	16	53797859	FTO	T	46.62%
rs1121980	16	53775335	FTO	T	42.13%

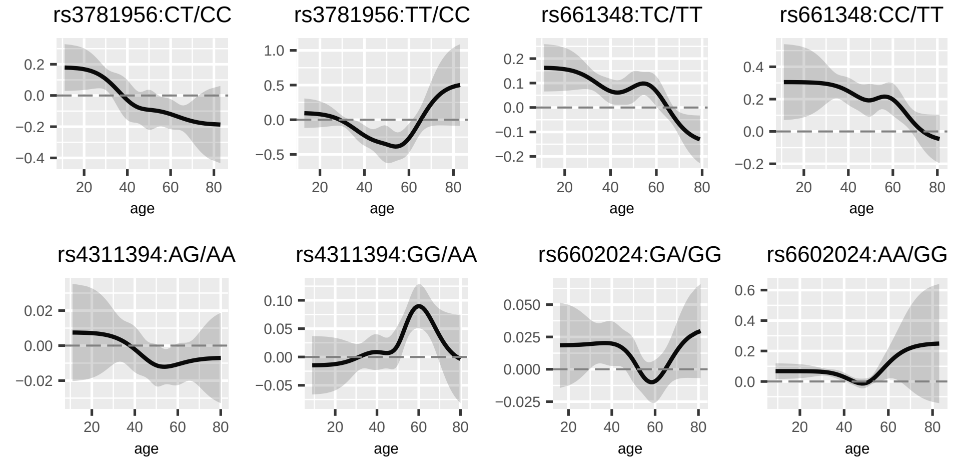

Fig 3: Compare the top 2 SNPs in our study and that in literature

References.

Imai, K. and Van Dyk, D. A. (2004). Causal inference with general treatment regimes. *Journal of the American Statistical Association* **99** 854–866.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79** 516–524.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* **39** 33–38.

Wanghuan Chu
Google Inc.
1600 Amphitheatre Parkway
Mountain View, CA, 94043, USA
E-mail: dqchuwh@gmail.com

Jingyuan Liu (Corresponding author)
MOE Key Laboratory of Econometrics
Department of Statistics, School of Economics
Wang Yanan Institute for Studies in Economics
and Fujian Key Lab of Statistics, Xiamen University
Xiamen, Fujian, 361005, China
E-mail: jingyuan@xmu.edu.cn

Runze Li
Department of Statistics
and the Methodology Center
Pennsylvania State University
State College, PA, 16801, USA
E-mail: rzli@psu.edu

Matthew Reimherr
Department of Statistics
Pennsylvania State University
State College, PA, 16801, USA
E-mail: mreimherr@psu.edu