Underestimation of SES effects in large cohorts: A DAG-informed simulation and NHANES case study

Ross A. Dunne

2025-08-14

```
<>:74: SyntaxWarning: invalid escape sequence '\*'
<>:74: SyntaxWarning: invalid escape sequence '\*'
/tmp/ipykernel_9311/713786669.py:74: SyntaxWarning: invalid escape sequence '\*'
  **Conclusions.** Across simulation and NHANES, **selection plus functional-form misspecificat
```

Background. Socioeconomic status (SES) is often modeled linearly or as coarse categories in large cohorts. When true SES→risk relationships are non-linear and samples are selected (healthy volunteer bias), SES's role can be under-estimated.

Methods. We built a DAG-informed simulation with latent (SES* affecting mediators (BMI, systolic BP, smoking) via non-linear functions and directly affecting a binary outcome. We generated a "biobank-like" sample by preferentially selecting higher (SES*) and lower risk. We compared typical models (linear/quintile SES; with/without mediator adjustment) to splines with g-computation of (E[Y mid do(SES=a)]). We summarized SES attribution via a causal variance share (R^2 causal) and a two-block Shapley split.

Results (simulation). Selection yielded a biobank fraction of **4.9%** ((N=9,851/200,000)) and reduced prevalence from **5.99%** (population) to **3.57%** (selected). In the selected sample, linear-quintile SES achieved **McFadden (R^2=0.011)**; including mediators raised predictive fit (**0.055**) while down-weighting SES as a putative cause. The SES causal (R^2) **changed by 88%** (ratio 1.88) from population to selected (oracle (SES^*)), and by **-91%** (ratio 0.09) when replacing oracle with a noisy proxy in the selected sample. In the NHANES case study using NHANES 2003–2018 (DEMO/BPX/BMX/MCQ/SMQ), spline + g-computation estimated an analog of SES causal (R^2) of **nan** (survey-weighted population) and **nan** (biobank-like selection), a **NA** change (ratio NA).

Conclusions. Across simulation and NHANES, **selection plus functional-form misspecification underestimates SES's contribution**. Flexible SES modeling (splines/GAMs) with standardization (or TMLE) recovers more of the causal signal and should be preferred in large cohorts.

```
None
```

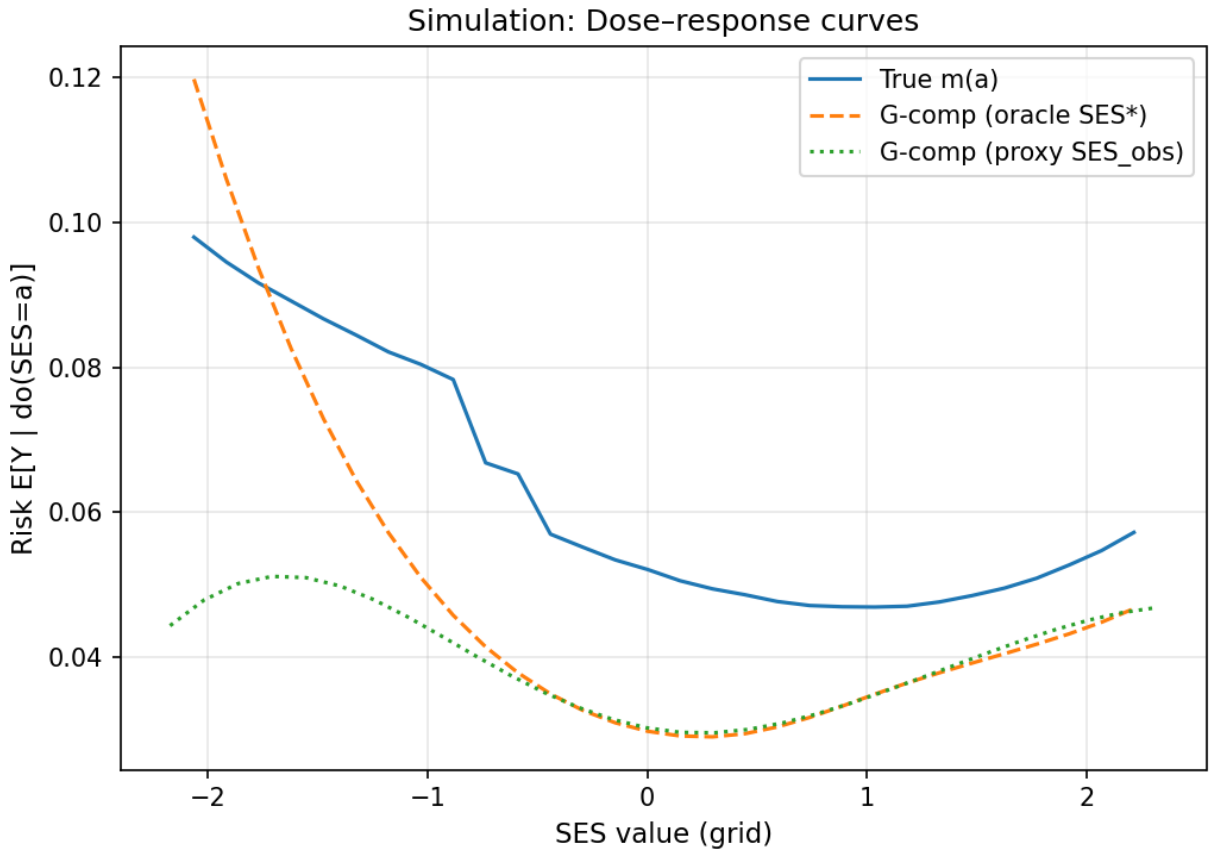

Figure 1: Simulation dose-response curves showing non-linear relationships and selection effects