Sensitivity Analysis of G-estimators to Invalid Instrumental Variables

Valentin Vancak, Arvid Sjölander

Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden

Table of Contents

Current section

Instrumental variable - Introduction

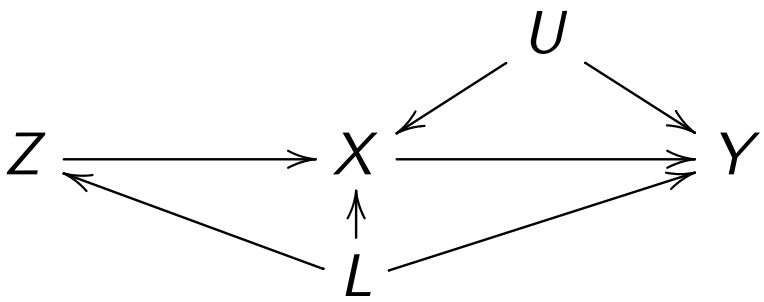
Instrumental variable

- Instrumental variable (IV) regression is a tool that is commonly used in analysis of observational data. Instrumental variables are used to make causal inference about the effect of a certain exposure in a presence of unmeasured confounders.
- A valid instrumental variable is a variable that is associated with the exposure, affects the outcome only through the exposure (exclusion criterion), and is unconfounded with the outcome (exogeneity).
- The IV assumptions are generally untestable and rely on subject-matter knowledge.
- We propose a new method of sensitivity analysis of the G-estimators to invalid instrumental variables. The new method is suitable for linear and non-linear models (specifically, logistic model), and requires only one sensitivity parameter both for the exclusion and the exogeneity assumptions.

Instrumental variable - Introduction

Directed acyclic graph (DAG) of a valid instrumental variable

- Let Y be the outcome variable, X the exposure, and Z the instrument. U represents all unmeasured confounders of X and Y, whereas L represents all measured confounders.
- The instrument Z affects Y only through the exposure X, and is unaffected by the unmeasured variables U.

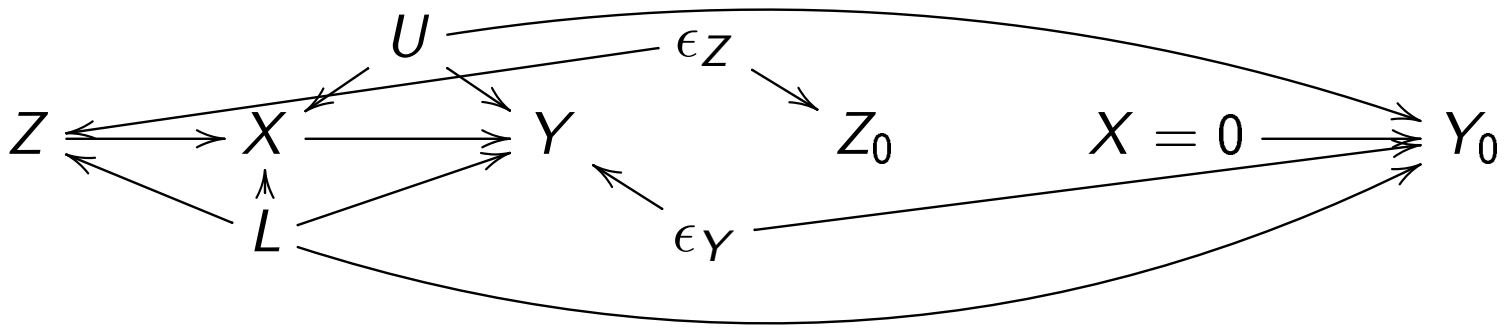


DAG of a causal model with a valid instrumental variable.

Instrumental variable - Introduction

Counterfactuals and the twin causal network

- A counterfactual implication of the exogeneity and exclusion assumptions is that a valid IV Z satisfies $Y_0 \perp\!\!\!\perp Z|L$.
- If Z is a valid IV, there is no open path between Y_0 and Z, conditionally on L.



DAG of a twin network causal model with a valid instrumental variable. The left-hand side of the DAG represents the actual world, while the right-hand side represents the hypothetical potential world. $X = 0$ represents the exposure X that is set to 0, and Z_0 represents the potential value of Z in this hypothetical setting.

Table of Contents

Causal model

Mean causal model

- Let Y_x be the potential outcome of Y when the exposure X is set to x. A general causal mean model is

$$\xi(E[Y_x|L,Z,X=x]) - \xi(E[Y_0|L,Z,X=x]) = m^T(L)x\psi,$$

  where ξ is a link-function of a generalized linear model, ψ is the vector of causal parameters, and $\dim(m(L)) = \dim(\psi)$.
- The composition of $m(L)$ defines the exact form of the causal model.
- Since $E[Y_x|L,Z,X=x] = E[Y|L,Z,X=x]$ this part is identifiable from the observed model.
- $E[Y_0|L,Z,X=x]$ is counterfactual, therefore its identification relies on the availability of a valid IV.

Causal model

Linear causal model

- For example, in a linear model, ξ is the identity link function. Assuming $m(L) = 1$, and a binary exposure X, the mean causal model is

$$E[Y_1|L, Z, X = 1] - E[Y_0|L, Z, X = 1] = \psi.$$

- Here ψ is the average causal effect of the exposure on the exposed.
- In linear models, ψ can be estimated using the two-stage least squares (TSLS) method. However, for non-linear models, the TSLS will produce inconsistent estimators.

Table of Contents

The G-estimator - Introduction

- The G-estimator is the value of the causal parameter ψ under which the assumption $Y_0 \perp\!\!\!\perp Z|L$ holds.
- The potential outcome Y_0 is counterfactual, whose mean is estimated by $h(\psi)$.
- The exact form of $h(\psi)$ depends on the link-function ξ

$$h(\psi) = \begin{cases} Y - m^T(L)X\psi, & \text{if } \xi \text{ is the identity function,} \\ Y exp\{-m^T(L)X\psi\}, & \text{if } \xi \text{ is the log link function,} \\ \text{expit}\left(\text{logit}(E[Y|X,Z,L] - m^T(L)X\psi\right), & \text{if } \xi \text{ is the logit function} \end{cases}$$

The G-estimator

- The G-estimator is obtained as a solution to a system of estimating equations. In particular, the G-estimator of a causal parameter ψ solves the following equation

$$\sum_{i=1}^{n} D(L_i, Z_i) h_i(\psi) = 0,$$

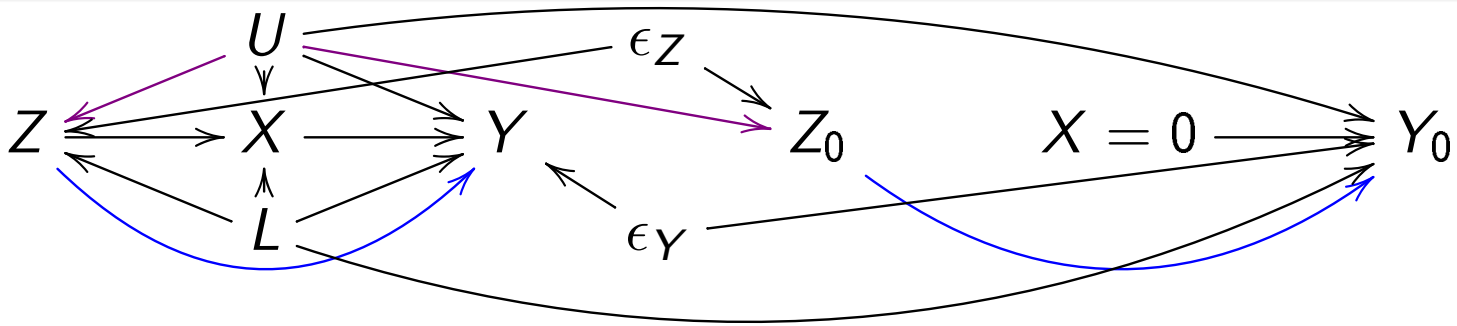
where $E[D(L, Z)|L] = 0$, and one of the common choices for D is

$$D(L, Z) = m(L)(Z - E[Z|L]).$$

- The consistency of the G-estimator relies on the validity of the instrument Z.
- The violation of the conditional independence of Y_0 and Z can arise either from the violation of the exclusion, the exogeneity assumption, or both. To illustrate this claim, we use the twin causal network.

Twin causal network with an invalid instrument

- The blue arrow from Z to Y represents the exclusion criterion violation, i.e., a direct effect of the instrument on the outcome. The violet arrow from U to Z represents represents the exogeneity assumption violation. $X = 0$ represents the exposure X that is set to 0, and Z_0 represents the potential value of Z in this setting.
- Since Z is not influenced by X, thus $Z_0 = Z$, therefore the assumption $Y_0 \perp\!\!\!\perp Z|L$ is violated by $Z_0 \to Y_0$,



DAG of a twin network causal model with invalid instrumental variable. The left-hand side of the DAG represents the actual world, while the right-hand side represents the hypothetical potential world.

Table of Contents

Sensitivity analysis of the G-estimators

Sensitivity parameter α

- We model compositions of mean independence violations using a parametric function $b(L, Z; \alpha)$, such that

$$\xi(E[Y_0|L, Z]) = a(L) + b(L, Z; \alpha).$$

  In the function $b(L, Z; \alpha)$, the parameter α is a sensitivity parameter that incorporates the conditional association between Y_0 and Z caused by the violation(s). Particularly, we require $b(L, Z; 0) = 0$, and $b(L, Z; \alpha) \neq 0$ for $\alpha \neq 0$.
- For any nonzero value of α, the G-estimator is inconsistent.

Sensitivity analysis of the G-estimators

Consistency of the G-estimators for invalid IVs

- To ensure the consistency of the G-estimators, we reformulate $h(\psi)$ as a function of α

$$h(\psi;\alpha) = \begin{cases} Y - m^T(L)X\psi - b(L,Z;\alpha), \\ Y\exp\{-m^T(L)X\psi - b(L,Z;\alpha)\}, \\ \text{expit}\left(\text{logit}(E[Y|X,Z,L] - m^T(L)X\psi - b(L,Z;\alpha)\right). \end{cases}$$

- For the true α, α^*, in $h(\psi;\alpha)$, the G-estimator is consistent.
- The true parameter α^* is non-identifiable and in real-world applications is rarely known.
- A sensitivity analysis is carried out by varying α over a range of plausible values, and mapping each value to a corresponding G-estimator.

Table of Contents

Logistic regression example

Causal parameters and violation on the logit scale

- Assume a binary outcome Y, a binary exposure X, a binary instrument Z, and an unmeasured confounder U. In addition, assume that there are no measured confounders, i.e., $L = \emptyset$, and $m(L) = 1$. We define the mean causal model on the logit scale

$$\text{logit}P(Y_1 = 1|X = 1, Z) - \text{logit}P(Y_0 = 1|X = 1, Z) = \psi.$$

Therefore, we define the violation on the logit scale as well

$$b(Z, L; \alpha^*) = \text{logit}P(Y_0 = 1|Z) - \text{logit}P(Y_0 = 1|Z = 0) = \alpha^* Z,$$

where α^* is the true violation parameter.

Logistic regression example

- Assuming logistic saturated outcome "model" for $E[Y|Z,X;\beta_Y]$ with an interaction term, such that $\beta_Y^T = (\beta_0, \beta_X, \beta_Z, \beta_{XZ})$. Therefore, $S(Y,X;\beta_Y)$ are the score functions of this model, particularly

$$S(Y,X;\beta_Y) = \begin{pmatrix} (Y - E[Y|Z,X;\beta_Y]) \\ (Y - E[Y|Z,X;\beta_Y])X \\ (Y - E[Y|Z,X;\beta_Y])Z \\ (Y - E[Y|Z,X;\beta_Y])XZ \end{pmatrix}.$$

- Let $D(L,Z;\mu_Z) = Z - \mu_Z$. Therefore, the G-estimator is the value of ψ that solves

$$\sum_{i=1}^{n}(Z_i - \mu_Z)\text{expit}\left(\text{logit}\hat{P}(Y_i = 1|Z_i, X_i; \hat{\beta}_Y) - X_i\psi - \alpha Z_i\right) = 0.$$

Logistic regression example

The system of estimating equations

- Let $\sum_{i=1}^{n} Q(Y_i, X_i, L_i, Z_i; \theta, \alpha) = 0$ be the estimating equations of the logistic causal model, where the vector of estimands is $\theta^T = (\beta_Y, \mu_Z, \psi)$, and

$$Q(Y, X, L, Z; \theta, \alpha) = \begin{pmatrix} S(Y, X, L; \beta_Y) \\ S(L, Z; \mu_Z) \\ D(L, Z; \mu_Z) h(\psi; \alpha) \end{pmatrix}.$$

- Notably, the score function of the instrument model $S(L, Z; \mu_Z)$ equals $D(L, Z; \mu_z)$.
- This is a system of unbiased estimating equations, i.e., $E[Q(Y, X, L, Z; \theta, \alpha)] = 0$, therefore the estimators that solve this system are consistent estimators of the true θ_0.

Logistic regression example

Asymptotic variance and distribution of the G-estimator

- The asymptotic variance of $\hat{\theta}$ is given by the sandwich formula

$$V(\theta_0, \alpha) = n^{-1}\mathrm{A}(\theta_0, \alpha)^{-1}\mathrm{B}(\theta_0, \alpha)\mathrm{A}(\theta_0, \alpha)^{-T}$$

  where $\mathrm{A}(\theta_0, \alpha) = E[-\partial Q(\theta_0, \alpha)/\partial\theta^T]$, $\mathrm{B}(\theta_0, \alpha) = E[Q(\theta_0, \alpha)Q(\theta_0, \alpha)^T]$, and θ_0 is the true value of the unknown parameters.

- The asymptotic distribution of the estimators $\hat{\theta}(\alpha)$ is multivariate normal, namely

$$\sqrt{n}(\hat{\theta}(\alpha) - \theta_0(\alpha)) \xrightarrow{D} \mathcal{N}_p(0, V(\theta_0, \alpha)).$$

- The asymptotic variance and the asymptotic distribution are general results, and do not apply only to the logistic model.

Logistic regression example - simulation

Logistic regression simulation - data generating process (DGP)

- The DGP is

$$
\begin{aligned}
Z &\sim Ber(p_z) \\
X|Z = z &\sim Ber\left(\text{expit}(\gamma_0 + \gamma_z z)\right) \\
Y|X = x, Z = z &\sim Ber\left(\text{expit}(\beta_0 + \beta_x x + \beta_z z + \beta_{xz} xz)\right).
\end{aligned}
$$

- We specify the marginal distribution of Y, X, and Z. In order to relate the violation structure to the DGP parameters of the observed data we use the fact that a valid IV satisfies $Y_0 \perp\!\!\!\perp Z|L$, particularly, $P(Y_0|Z = 1) = P(Y_0|Z = 0)$ assuming that $L = \emptyset$. Therefore, for an invalid IV, the violation structure is

$$
\text{logit}P(Y_0|Z = 1) - \text{logit}P(Y_0|Z = 0) = \alpha^*.
$$

Logistic regression example - simulation

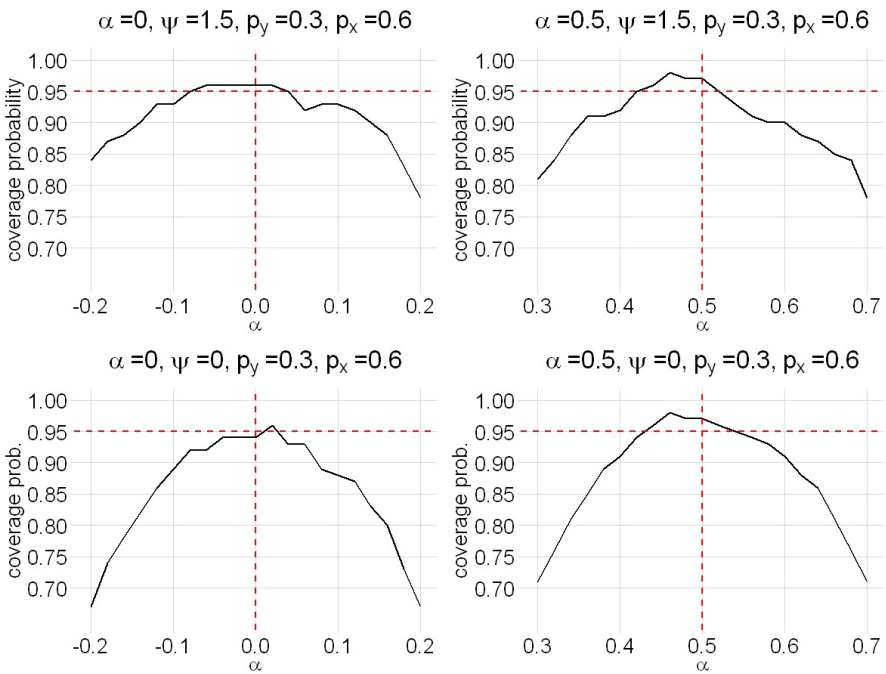
Logistic regression simulation - data generating process (DGP)

- By using the causal parameter ψ and the DGP we obtain a function w.r.t. the unknown parameters of the observed data

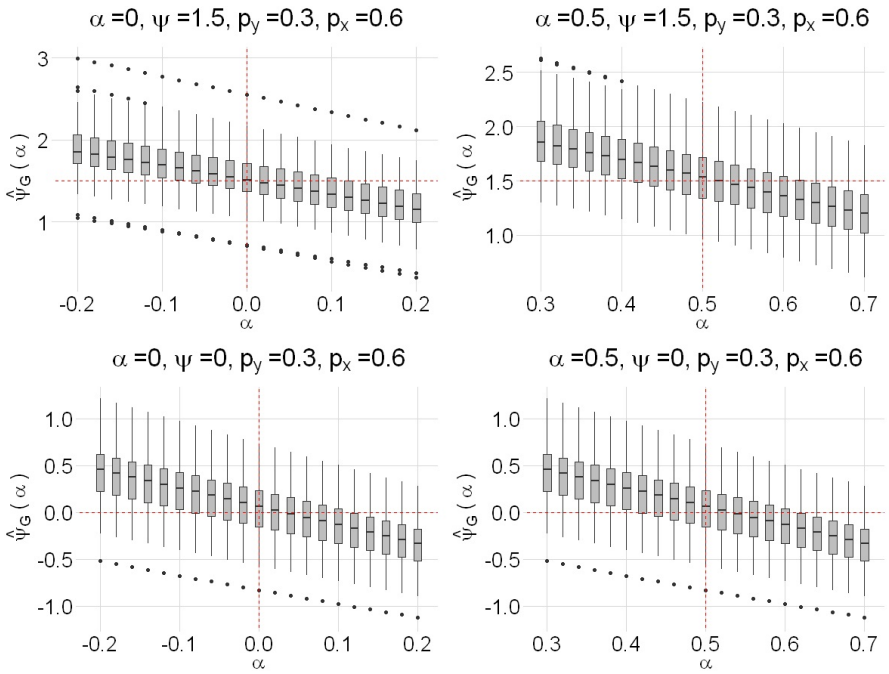
$$
\begin{aligned}
&P(Y_0 = 1|Z) \\
&= \text{expit}(\beta_0 + \beta_z)(1 - \text{expit}(\gamma_0 + \gamma_z)) \\
&+ \text{expit}(\beta_0 + \beta_x + \beta_z + \beta_{xz} - \psi)\text{expit}(\gamma_0 + \gamma_z)\,.
\end{aligned}
$$

- By plugging in this result in the equation that defines the violation structure, we obtain the functional relationship between the true sensitivity parameter α^* and the DGP parameters of the observed data.
- The parameters of the DGP are determined after the specification of the true ψ and $\alpha*$

Logistic regression example - simulation results

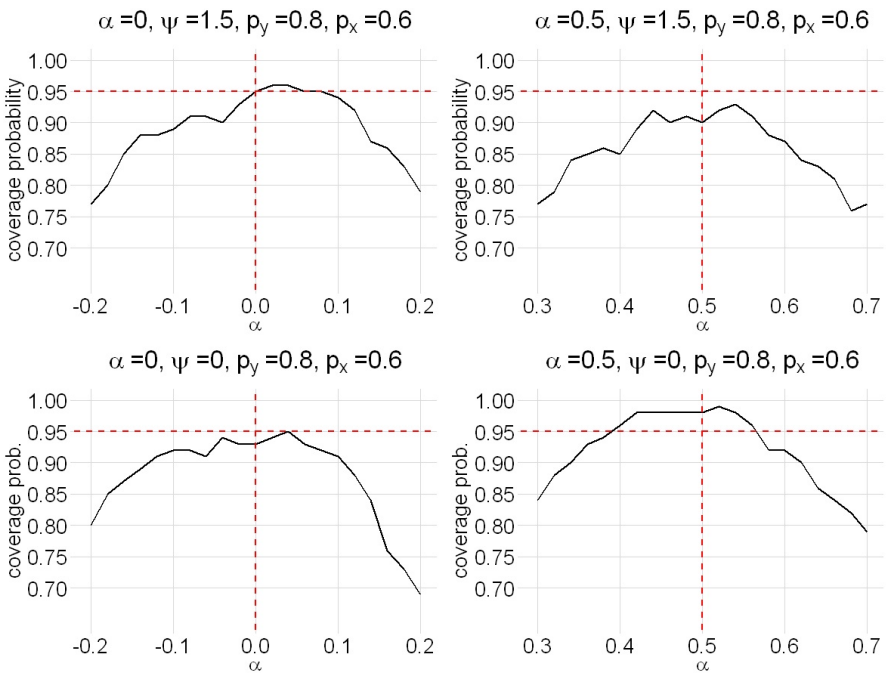


Coverage rates of the 95% CIs of the true causal parameter $\psi \in \{0, 1.5\}$ in the logistic causal model as a function of the sensitivity parameter α, for $\alpha^* \in \{0, 0.5\}$, for sample size $n = 1000$, and $m = 100$ repetitions. Additionally, $P(Y = 1) = p_y = 0.3$.

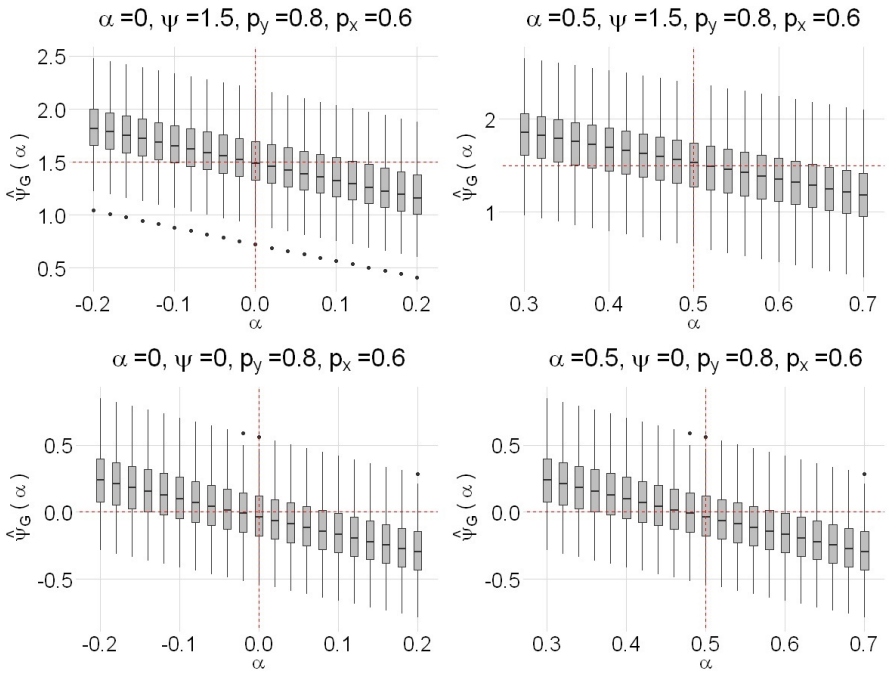


Boxplots of empirical distribution of $\hat{\psi}_G(\alpha)$ for the true causal parameter $\psi \in \{0, 1.5\}$ as a function of the sensitivity parameter α, for $\alpha^* \in \{0, 0.5\}$, for sample size $n = 1000$, and $m = 100$ repetitions. Additionally, $P(Y = 1) = p_y = 0.3$.

Logistic regression example - simulation results



Coverage rates of the 95% CIs of the true causal parameter $\psi \in \{0, 1.5\}$ in the logistic causal model as a function of the sensitivity parameter α, for $\alpha^* \in \{0, 0.5\}$, for sample size $n = 1000$, and $m = 100$ repetitions. Additionally, $P(Y = 1) = p_y = 0.8$.



Boxplots of empirical distribution of $\hat{\psi}_G(\alpha)$ for the true causal parameter $\psi \in \{0, 1.5\}$ as a function of the sensitivity parameter α, for $\alpha^* \in \{0, 0.5\}$, for sample size $n = 1000$, and $m = 100$ repetitions. Additionally, $P(Y = 1) = p_y = 0.8$.

Logistic regression example - simulation summary

Simulation summary

- For the true value of α in $h(\psi;\alpha)$, the G-estimators are consistent and their asymptotic confidence intervals meet the nominal coverage probabilities.
- The proposed method works well both for valid and invalid IVs, and for linear and logistic models.
- The proposed method works well also when the true causal effect ψ is zero, both for linear and logistic models.
- In the logistic causal model with logistic outcome model, the variance of the G-estimator depends on the dimension and the stability of the parameters estimators of the outcome model. Namely, compared to the linear causal models, the CIs are wider, and therefore, the coverage probability of the CI depends less on the assumed value of α.

Table of Contents

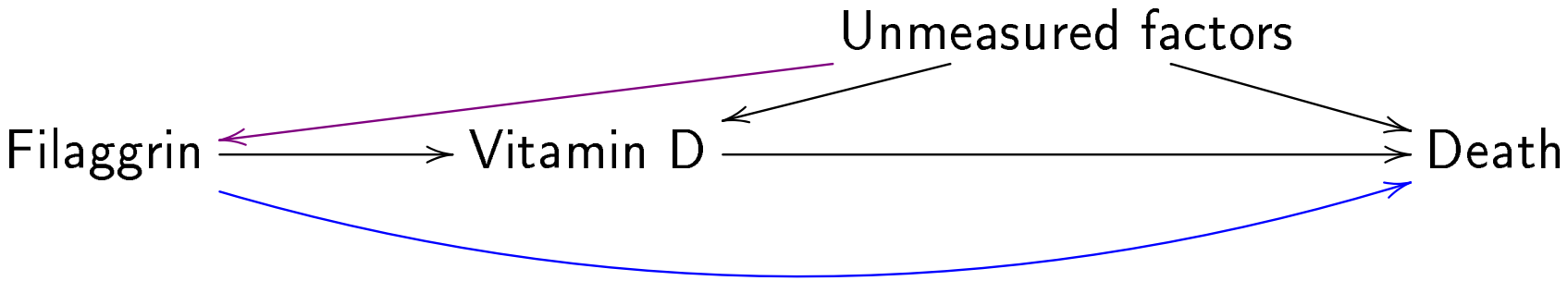
Real world example - effects of vitamin D on mortality

- Vitamin D deficiency has been linked with several lethal conditions such as cancer and cardiovascular diseases.
- Vitamin D status is also associated with unmeasured behavioral and environmental factors that may result in biased estimators when using standard statistical analyses to estimate causal effects.
- Mutations in the filaggrin gene have been shown to be associated with a higher vitamin D status and are assumed to satisfy the IV assumptions.
- We use a publicly available mutilated version of the data on a cohort study on vitamin D status causal effect on mortality rates.
- The data frame contains 2571 subjects and 5 variables: age (at baseline), filaggrin (a binary indicator of whether filaggrin mutations are present), vitd (vitamin D level at baseline, measured as serum 25-OH-D(nmol/L)), time (follow-up time), and death (an indicator of whether the subject died during follow-up).

Real world example - effects of vitamin D on mortality

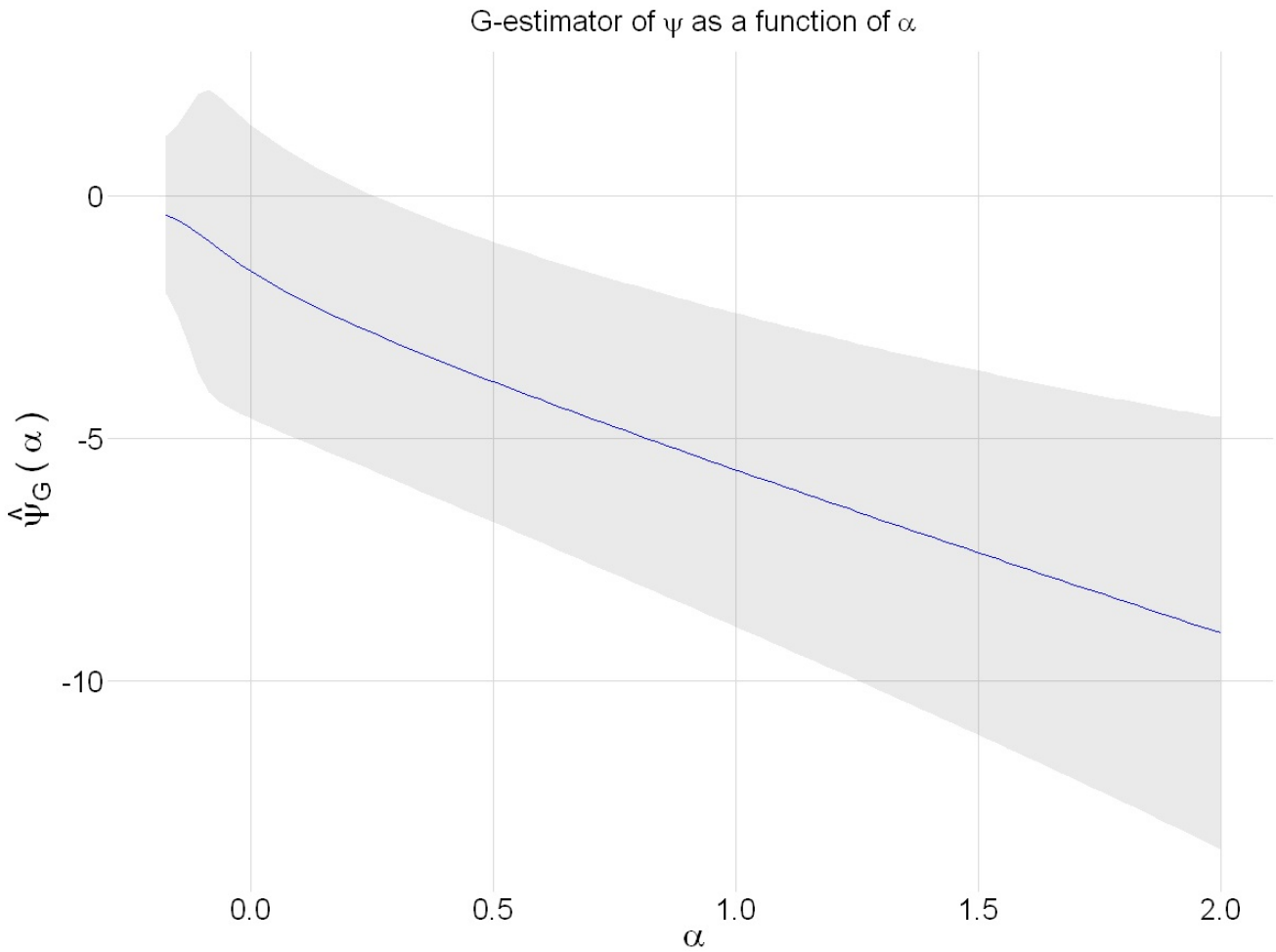
Vitamin D data

- The death during follow-up is the point outcome Y.
- The presence of the filaggrin gene mutations is the IV Z.
- Following Martinussen et al. (2019) the scaled version of the vitamin D status at baseline is a continuous exposure variable X.



DAG of the vitamin D status model. The violet arrow represents a possible exogeneity assumption violation, and the blue line represents a possible exclusion assumption violation.

Real world example - effects of vitamin D status on mortality



G-estimator $\hat{\psi}_G(\alpha)$ of the vit. D status causal effect on death rate during follow-up as a function of the sensitivity parameter α. For $\alpha = 0$ the $\hat{\psi}_G(0) = -1.558$. For $\alpha < -0.17$ there is no solution to the estimating equations.

Table of Contents

Conclusions

Vitamin D data

- If the mean models are correctly specified, there is no evidence of a positive causal effect of vitamin D deficiency on the mortality rate.
- If the IV is invalid, then the true causal effect is likely to be of larger magnitude than the estimated value for $\alpha = 0$.

Study

- This study provides theoretical framework and practical guidelines on how to conduct sensitivity analysis of G-estimators using single sensitivity parameter that captures violations of both the exogeneity and the exclusion assumptions.
- The proposed method is applicable both to linear and non-linear causal models.

Thank you!