Supplementary material for "Average causal effect estimation via instrumental variables: the no simultaneous heterogeneity assumption" by Hartwig FP, Wang L, Davey Smith G and Davies NM

Table of contents

1. Non-parametric structural equation models (SEMs)

In this section we describe the SEMs corresponding to the causal graphs and assumptions required for NOSH to hold. Initially, the SEM corresponding to the unrestricted model (depicted in Figure 2A) is:

$$U_{1_i} = f_{U_1}\left(Z = Z_i, \varepsilon_L = \varepsilon_{L_i}, \varepsilon_{U_1} = \varepsilon_{U_{1_i}}\right) \quad \text{(S1.1)}$$

$$U_{2_i} = f_{U_2}\left(X = X_i, \varepsilon_L = \varepsilon_{L_i}, \varepsilon_{U_2} = \varepsilon_{U_{2_i}}\right) \quad \text{(S2.1)}$$

$$U_{3_i} = f_{U_3}\left(\varepsilon_L = \varepsilon_{L_i}, \varepsilon_{U_3} = \varepsilon_{U_{3_i}}\right) \quad \text{(S3.1)}$$

$$U_{4_i} = f_{U_4}\left(\varepsilon_L = \varepsilon_{L_i}, \varepsilon_{U_4} = \varepsilon_{U_{4_i}}\right) \quad \text{(S4.1)}$$

$$U_{5_i} = f_{U_5}\left(\varepsilon_L = \varepsilon_{L_i}, \varepsilon_{U_5} = \varepsilon_{U_{5_i}}\right) \quad \text{(S5.1)}$$

$$U_{6_i} = f_{U_6}\left(\varepsilon_L = \varepsilon_{L_i}, \varepsilon_{U_6} = \varepsilon_{U_{6_i}}\right) \quad \text{(S6.1)}$$

$$Z_i = f_Z\left(\varepsilon_Z = \varepsilon_{Z_i}\right) \quad \text{(S7.1)}$$

$$X_i = f_X\left(Z = Z_i, U_1 = U_{1_i}, U_3 = U_{3_i}, U_4 = U_{4_i}, U_5 = U_{5_i}, U_6 = U_{6_i}, \varepsilon_X = \varepsilon_{X_i}\right) \quad \text{(S8.1)}$$

$$\beta_{X_i} = f_{\beta_X}\left(Z = Z_i, U_1 = U_{1_i}, U_4 = U_{4_i}, U_6 = U_{6_i}, \varepsilon_{\beta_X} = \varepsilon_{\beta_{X_i}}\right) \quad \text{(S9.1)}$$

$$\beta_{Y_i} = f_{\beta_Y}\left(X = X_i, U_2 = U_{2_i}, U_5 = U_{5_i}, U_6 = U_{6_i}, \varepsilon_{\beta_Y} = \varepsilon_{\beta_{Y_i}}\right) \quad \text{(S10.1)}$$

In equations (S1.1)-(S6.1), ε_L denotes the latent variable that is a cause of all unmeasured variables, thus rendering all of them d-connected to one another.

Since d-separation implies statistical independence, Figure 2 and equations S1.1-S10.1 can be used to find conditions under which β_Y is d-separated to β_X and Z, and thus statistically independent. From equations S9.1 and S10.1, it is clear that β_X and β_Y are d-connected due to shared terms (i.e., U_6), terms that may be correlated (e.g., U_4 and U_5) and terms that are functionally related (e.g., X_i and Z_i) in the r.h.s. of these equations. Moreover, since equation S10.1 contains X in the r.h.s., β_Y and Z are clearly d-connected.

Given Assumption 1, equations S1.1-S10.1 become:

$$U_{1_i} = f_{U_1}\left(Z = Z_i, \varepsilon_{L_1} = \varepsilon_{L_{1_i}}, \varepsilon_{U_1} = \varepsilon_{U_{1_i}}\right) \quad \text{(S1.2)}$$

$$U_{2_i} = f_{U_2}\left(\varepsilon_{L_2} = \varepsilon_{L_{2_i}}, \varepsilon_{U_2} = \varepsilon_{U_{2_i}}\right) \quad \text{(S2.2)}$$

$$U_{3_i} = f_{U_3}\left(\varepsilon_{L_1} = \varepsilon_{L_{1_i}}, \varepsilon_{L_2} = \varepsilon_{L_{2_i}}, \varepsilon_{U_3} = \varepsilon_{U_{3_i}}\right) \quad \text{(S3.2)}$$

$$U_{4_i} = f_{U_4}\left(\varepsilon_{L_1} = \varepsilon_{L_{1_i}}, \varepsilon_{U_4} = \varepsilon_{U_{4_i}}\right) \quad \text{(S4.2)}$$

$$U_{5_i} = f_{U_5}\left(\varepsilon_{L_2} = \varepsilon_{L_{2_i}}, \varepsilon_{U_5} = \varepsilon_{U_{5_i}}\right) \quad \text{(S5.2)}$$

$$U_{6_i} = \emptyset \quad \text{(S6.2)}$$

$$Z_i = f_Z\left(\varepsilon_Z = \varepsilon_{Z_i}\right) \quad \text{(S7.2)}$$

$$X_i = f_X\left(Z = Z_i, U_1 = U_{1_i}, U_3 = U_{3_i}, U_4 = U_{4_i}, U_5 = U_{5_i}, \varepsilon_X = \varepsilon_{X_i}\right) \quad \text{(S8.2)}$$

$$\beta_{X_i} = f_{\beta_X}\left(Z = Z_i, U_1 = U_{1_i}, U_4 = U_{4_i}, \varepsilon_{\beta_X} = \varepsilon_{\beta_{X_i}}\right) \quad \text{(S9.2)}$$

$$\beta_{Y_i} = f_{\beta_Y}\left(X = X_i, U_2 = U_{2_i}, U_5 = U_{5_i}, \varepsilon_{\beta_Y} = \varepsilon_{\beta_{Y_i}}\right) \quad \text{(S10.2)},$$

where ε_{L_1} and ε_{L_2} such that $\varepsilon_{L_1} \perp\!\!\!\perp \varepsilon_{L_2}$ are latent variables that allow U_1, U_3 and U_4 to be correlated with one another and U_2, U_3 and U_5 to be correlated with one another.

Assumption 2 implies that X is not in the r.h.s. of equation S10.2, which becomes:

$$\beta_{Y_i} = f_{\beta_Y}\left(U_2 = U_{2_i}, U_5 = U_{5_i}, \varepsilon_{\beta_Y} = \varepsilon_{\beta_{Y_i}}\right) \quad \text{(S10.3)}.$$

2. Implications of non-linear effects and data-generating models for NOSH

Under linear models for X and Y, it is worth emphasizing the role of non-linear effects for NOSH. As can be seen in Figure 2A, if both the association of Z on X and the effect of X on Y are non-linear on the additive scale, the following path exists: $\beta_X \leftarrow Z \rightarrow X \rightarrow \beta_Y$, implying that NOSH is violated. That is, when both the association of Z and X and the effect of X on Y are non-linear, NOSH is violated regardless of unmeasured effect modifiers due to violations of Assumption 2.

We now discuss implications of multiplicative models. For simplicity of exposition, assume that Z, X, Y, V_X and V_Y (with V_X and V_Y denoting unmeasured variables) are all binary. $X_i \sim \text{Bernoulli}\left(\pi_{X_i}\right)$, where $0 \leq \pi_{X_i} = \omega_X \times \alpha^{Z_i} \times \tau^{V_{X_i}} \leq 1$. This model can be re-written as $\ln(\text{E}[X|Z, V_X]) = \ln(\omega_X) +$

$Z\ln(\alpha) + V_X\ln(\theta)$, which clearly shows there is no multiplicative effect modification. This could be interpreted as $V_X \in U_3$. However, lack of multiplicative effect modification implies additive effect modification, which is the relevant scale for β_X. In this example, this happens because:

$$\mathrm{E}[X|Z=1,V_X=0] - \mathrm{E}[X|Z=0,V_X=0] = \omega_X\alpha - \omega_X = \omega_X(\alpha-1)$$

$$\mathrm{E}[X|Z=1,V_X=1] - \mathrm{E}[X|Z=0,V_X=1] = \omega_X\alpha\tau - \omega_X\tau = \omega_X\tau(\alpha-1)$$

The above illustrates that, under a multiplicative model on X, β_X will likely vary according to other causes of X (in this case, V_X), unless in rather contrived scenarios the model parameters cancel one another with respect to additive effect modification (e.g., if multiplicative effect modification leads exactly to no additive effect modification). Graphically, this implies that the causal diagram should contain an arrow from V_X to β_X under a multiplicative model on X. Therefore, $V_X \notin U_3$, being a potential member of U_1, U_4 or U_6.

Now suppose that $Y_i \sim \mathrm{Bernoulli}(\pi_{Y_i})$, where $0 \leq \pi_{Y_i} = \mu_Y \times \rho^{X_i} \times \chi^{V_{Y_i}} \leq 1$. Again, there is no multiplicative effect modification, but there is additive effect modification:

$$\mathrm{E}[Y|X=1,V_Y=0] - \mathrm{E}[Y|X=0,V_Y=0] = \mu_Y\rho - \mu_Y = \mu_Y(\rho-1)$$

$$\mathrm{E}[Y|X=1,V_Y=1] - \mathrm{E}[Y|X=0,V_Y=1] = \mu_Y\rho\chi - \mu_Y\chi = \mu_Y\chi(\rho-1)$$

The above illustrates that, under a multiplicative model on Y, β_Y will likely vary according to other causes of Y (in this case, V_Y). Graphically, this implies that the causal diagram should contain an arrow from V_X to β_X. Therefore, $V_X \notin U_3$, being a potential member of U_2, U_5 or U_6.

This simple example illustrates the more general implication of the fact that additive effect modification is implied by lack of multiplicative effect modification for the plausibility of NOSH. Under a multiplicative model for X and a linear model for Y, any cause of X (here denoted by V_X) will generally be an effect modifier of β_X. Therefore, if V_X is also either an effect modifier or a surrogate effect modifier of β_Y, Assumption 1 is violated. This implies that, unless heterogeneity in β_Y is independent of all causes of X, Assumption 1 will likely be violated.

Under a linear model for X and a multiplicative model for Y, any cause of Y (here denoted by V_Y) will generally by an effect modifier of β_Y. Therefore, if V_Y is also either an effect modifier or a surrogate effect modifier of β_X, then Assumption 1 would be violated. This implies that, unless heterogeneity in β_X is independent of all causes of Y, Assumption 1 would likely be violated.

Under multiplicative models for both X and Y, both V_X and V_Y will generally be modifiers of β_X and β_Y, respectively. This implies that, if V_X and V_Y are statistically dependent, then both will likely be either effect modifiers or surrogate effect modifiers of both β_X and β_Y, thus violating Assumption 1. This will happen, for example, if $V_X = V_Y$ or if these variables have a common cause (i.e., if there is unmeasured confounding). Since assuming no unmeasured confounding would eliminate the need of using instrumental variables, this is not a sensible assumption in this context.

3. Proof of theorem 1

We now prove Theorem 1. From the notation introduced in section 2.2~~.1~~ (recall that U_1 is not a cause of Y and U_3 and U_4 do not modify the causal effect of X on Y), the non-parametric structural equation model governing Y can be equivalently defined as $Y_i = h(X_i, U_{2i}, U_{5i}, U_{6i}, \varepsilon_{Yi}) + g(U_{2i}, U_{3i}, U_{4i}, U_{5i}, U_{6i}, \varepsilon_{Yi})$, where $h(.)$ is a generic function such that that all of its terms include X, and $g(.)$ is another generic function that does not include X as one of its arguments. From this, the function $F_{Yi}(x)$ defined in section 2.1 can be equivalently defined as $F_{Yi}(X_i) = \mathrm{E}[h(X_i, U_{2i}, U_{5i}, U_{6i}, \varepsilon_{Yi}) + g(U_{2i}, U_{3i}, U_{4i}, U_{5i}, U_{6i}, \varepsilon_{Yi})|do(X_i = x), U = U_i, \varepsilon_Y = \varepsilon_{Yi}]$.

From this, the numerator of the Wald estimand ($\mathrm{cov}(Y_i, Z_i)$) can defined as follows:

$$\mathrm{E}[Y_i Z_i] = \mathrm{E}\left[\left(h(X_i, U_{2i}, U_{5i}, U_{6i}, \varepsilon_{Yi}) + g(U_{2i}, U_{3i}, U_{4i}, U_{5i}, U_{6i}, \varepsilon_{Yi})\right) Z_i\right]$$

$$= \mathrm{E}[h(X_i, U_{2i}, U_{5i}, U_{6i}, \varepsilon_{Yi}) Z_i] + \mathrm{E}[g(U_{2i}, U_{3i}, U_{4i}, U_{5i}, U_{6i}, \varepsilon_{Yi})]\mathrm{E}[Z_i],$$

where $\mathrm{E}[g(U_{2i}, U_{3i}, U_{4i}, U_{5i}, U_{6i}, \varepsilon_{Yi}) Z_i] = \mathrm{E}[g(U_{2i}, U_{3i}, U_{4i}, U_{5i}, U_{6i}, \varepsilon_{Yi})]\mathrm{E}[Z_i]$ holds given the independence assumption (section 2.1).

$$\mathrm{E}[Y_i]\mathrm{E}[Z_i] = \mathrm{E}[h(X_i, U_{2i}, U_{5i}, U_{6i}, \varepsilon_{Yi}) + g(U_{2i}, U_{3i}, U_{4i}, U_{5i}, U_{6i}, \varepsilon_{Yi})]\mathrm{E}[Z_i]$$

$$= \mathrm{E}[h(X_i, U_{2i}, U_{5i}, U_{6i}, \varepsilon_{Yi})]\mathrm{E}[Z_i] + \mathrm{E}[h_2(U_{2i}, U_{3i}, U_{4i}, U_{5i}, U_{6i}, \varepsilon_{Yi})]\mathrm{E}[Z_i]$$

$$\mathrm{cov}(Y_i, Z_i) = \mathrm{E}[Y_i Z_i] - \mathrm{E}[Y_i]\mathrm{E}[Z_i]$$

$$= \mathrm{E}[h(X_i, U_{2i}, U_{5i}, U_{6i}, \varepsilon_{Yi}) Z_i] - \mathrm{E}[h(X_i, U_{2i}, U_{5i}, U_{6i}, \varepsilon_{Yi})]\mathrm{E}[Z_i]$$

$$= \mathrm{cov}(h(X_i, U_{2i}, U_{5i}, U_{6i}, \varepsilon_{Yi}), Z_i)$$

$$= \mathrm{cov}(\mathrm{E}[h(X_i, U_{2i}, U_{5i}, U_{6i}, \varepsilon_{Yi})|X = X_i, U = U_i, \varepsilon_Y = \varepsilon_{Yi}], Z_i).$$

The last equality follows from $\text{cov}(Y_i, Z_i) = \text{cov}\big(F_{Y_i}(X_i), Z_i\big)$ and that, for individual i, conditioning on $do(X = X_i)$ is equivalent to conditioning on the observed value of X under the stable unit treatment value assumption.

Under Assumption 1, $U_6 = \emptyset$. Under Assumption 2, $\text{E}\big[h\big(X_i, U_{2i}, U_{5i}, U_{6i}, \varepsilon_{Y_i}\big)|X = X_i, U = U_i, \varepsilon_Y = \varepsilon_{Y_i}\big] = \text{E}\big[h^{\#}\big(U_{2i}, U_{5i}, U_{6i}, \varepsilon_{Y_i}\big)|U = U_i, \varepsilon_Y = \varepsilon_{Y_i}\big]X_i$, where $h^{\#}(.)$ is a generic function. In this case, $F_{Y_i}(1) - F_{Y_i}(0) = \text{E}\big[h^{\#}\big(U_{2i}, U_{5i}, U_{6i}, \varepsilon_{Y_i}\big)|U = U_i, \varepsilon_Y = \varepsilon_{Y_i}\big]$ and (assuming differentiability with respect to X) $\frac{\partial}{\partial x}\big[F_{Y_i}(x)\big]\Big|_{x=X_i} = \text{E}\big[h^{\#}\big(U_{2i}, U_{5i}, U_{6i}, \varepsilon_{Y_i}\big)|U = U_i, \varepsilon_Y = \varepsilon_{Y_i}\big]$. That is, for either a binary or a continuous X, $\beta_{Y_i} = \text{E}\big[h^{\#}\big(U_{2i}, U_{5i}, U_{6i}, \varepsilon_{Y_i}\big)|U = U_i, \varepsilon_Y = \varepsilon_{Y_i}\big]$.

Therefore, if NOSH holds, $\text{cov}(Y_i, Z_i) = \text{cov}\big(\beta_{Y_i}X_i, Z_i\big) = \text{E}\big[\beta_{Y_i}\big]\text{cov}(X_i, Z_i)$. The last equality holds because:

$$\text{cov}\big(\beta_{Y_i}X_i, Z_i\big) = \text{cov}\big(\text{E}\big[\beta_{Y_i}X_i|Z_i\big], \text{E}\big[Z_i|\beta_{Y_i}\big]\big) + \text{E}\big[\text{cov}\big(\beta_{Y_i}X_i, Z_i|\beta_{Y_i}\big)\big]$$

$$= \text{E}\big[\text{cov}\big(\beta_{Y_i}X_i, Z_i|\beta_{Y_i}\big)\big].$$

This equality holds because $\text{E}\big[Z_i|\beta_{Y_i}\big] = \text{E}[Z_i]$, so that $\text{cov}\big(\text{E}\big[\beta_{Y_i}X_i|Z_i\big], \text{E}\big[Z_i|\beta_{Y_i}\big]\big) = \text{cov}\big(\text{E}\big[\beta_{Y_i}X_i|Z_i\big], \text{E}[Z_i]\big) = 0$. Therefore:

$$\text{cov}\big(\beta_{Y_i}X_i, Z_i\big) = \text{E}\big[\text{cov}\big(\beta_{Y_i}X_i, Z_i|\beta_{Y_i}\big)\big] = \text{E}\big[\beta_{Y_i}\text{cov}\big(X_i, Z_i|\beta_{Y_i}\big)\big]$$

$$= \text{E}\big[\beta_{Y_i}\text{cov}(X_i, Z_i)\big] = \text{E}\big[\beta_{Y_i}\big]\text{cov}(X_i, Z_i).$$

The equality $\text{E}\big[\beta_{Y_i}\text{cov}\big(X_i, Z_i|\beta_{Y_i}\big)\big] = \text{E}\big[\beta_{Y_i}\text{cov}(X_i, Z_i)\big]$ holds because, by NOSH, $\beta_Y \perp\!\!\!\perp \beta_X$.

From this, $\beta_{IV} = \frac{\text{E}[\beta_{Y_i}]\text{cov}(X_i, Z_i)}{\text{cov}(X_i, Z_i)} = \text{E}\big[\beta_{Y_i}\big] = \text{ACE}$, thus finishing the proof.

Although the above proves Theorem 1 in generality, it is instructive to consider alternative proofs under specific scenarios. We will also use these additional proofs to show that only mean independence is sufficient for identification. Initially, consider the case where Z and X are binary. In this case, $\beta_{X_i} \in \{1, -1, 0\}$, respectively indicating that individual i is a complier, a defier or either a never or always taker. In this case, β_{IV} can be defined as $\beta_{IV} = \frac{\text{E}[\beta_{Y_i}|\beta_{X_i}=1]P(\beta_{X_i}=1) - \text{E}[\beta_{Y_i}|\beta_{X_i}=-1]P(\beta_{X_i}=-1)}{P(\beta_{X_i}=1) - P(\beta_{X_i}=-1)}$, which is essentially a weighted average of conditional ACEs within subgroups where the relevance assumption hold, with the sign of the weights corresponding to the sign of the Z-X association.[1] Because X is binary, the effect of X on Y is

necessarily additive linear (i.e., Assumption 2 necessarily holds). If Assumption 1 also holds (even if β_Y and β_X are mean independent, but not fully independent), then $\mathrm{E}[\beta_{Y_i}|\beta_{X_i} = 1] = \mathrm{E}[\beta_{Y_i}|\beta_{X_i} = -1] = \mathrm{E}[\beta_{Y_i}] = \mathrm{ACE}$. Since β_{IV} is a weighted average of conditional effects, $\beta_{IV} = \mathrm{ACE}$.

Now, consider the case where Z is binary and X is continuous. In this case, the effects of Z on X and on Y are necessarily additive linear, implying that $\beta_{X_i} = X_i(Z_i = 1) - X_i(Z_i = 0) = \frac{\partial X_i}{\partial Z}$, where $X_i(Z_i = z) = f_X(do(Z = z), U = U_i, \varepsilon_X = \varepsilon_{X_i})$ for $z \in \{0,1\}$. Therefore, the denominator of the Wald estimand is $\mathrm{E}[\beta_{X_i}] = \mathrm{E}\left[\frac{\partial X_i}{\partial Z}\right]$. The numerator of the Wald estimand is: $\mathrm{E}[Y_i(X_i(Z_i = 1)) - Y_i(X_i(Z_i = 0))] = \mathrm{E}\left[\frac{(Y_i(X_i(Z_i=1)) - Y_i(X_i(Z_i=0)))}{X_i(Z_i=1) - X_i(Z_i=0)}(X_i(Z_i = 1) - X_i(Z_i = 0))\right] = \mathrm{E}\left[\frac{(Y_i(X_i(Z_i=1)) - Y_i(X_i(Z_i=0)))}{X_i(Z_i=1) - X_i(Z_i=0)}\frac{\partial X_i}{\partial Z}\right]$, where $Y_i(X_i(Z_i = z)) = f_Y(do(X = X_i(Z_i = z)), U = U_i, \varepsilon_Y = \varepsilon_{Y_i})$. Notice that, for simplicity, the definitions of $X_i(Z_i = z)$ and $Y_i(X_i(Z_i = z))$ assume deterministic counterfactuals. Therefore, the numerator of the Wald estimand equals $\mathrm{E}\left[\frac{\partial Y_i}{\partial X}\frac{\partial X_i}{\partial Z}\right]$ if the effect of X on Y is additive linear, which is true under NOSH, which also implies $\frac{\partial Y_i}{\partial X} \perp\!\!\!\perp \frac{\partial X_i}{\partial Z}$ (or just $\mathrm{cov}\left(\frac{\partial Y_i}{\partial X}, \frac{\partial X_i}{\partial Z}\right) = 0$ under the relaxed version of NOSH). In this case, $\beta_{IV} = \frac{\mathrm{E}\left[\frac{\partial Y_i}{\partial X}\frac{\partial X_i}{\partial Z}\right]}{\mathrm{E}\left[\frac{\partial X_i}{\partial Z}\right]} = \frac{\mathrm{E}\left[\frac{\partial Y_i}{\partial X}\right]\mathrm{E}\left[\frac{\partial X_i}{\partial Z}\right]}{\mathrm{E}\left[\frac{\partial X_i}{\partial Z}\right]} = \mathrm{E}\left[\frac{\partial Y_i}{\partial X}\right] = \mathrm{E}[\beta_{Y_i}]$.[2]

The case where Z is multivalued follows from the arguments above. For a multivalued discrete Z coded such that $\mathrm{E}[X|Z = 1] \leq \cdots \leq \mathrm{E}[X|Z = K]$, where K is the number of values that Z attains, let $\beta_{IV_{Z,Z-1}} = \frac{\mathrm{E}[Y_i|Z=z] - \mathrm{E}[Y_i|Z=z-1]}{\mathrm{E}[X_i|Z=z] - \mathrm{E}[X_i|Z=z-1]}$. The IV estimand is $\beta_{IV} = \sum_{z=2}^{K} \beta_{IV_{Z,Z-1}}\omega_z$, where ω_z denotes the weight that $\beta_{IV_{Z,Z-1}}$ receives in the overall estimand, defined such that $\sum_{z=2}^{K} \omega_z = 1$.[3,4] Notice that $\beta_{IV_{Z,Z-1}}$ is essentially the Wald estimand for a binary instrument but restricted to the subset $Z \in \{z - 1, z\}$. Therefore, by the arguments above, $\beta_{IV_{Z,Z-1}} = \mathrm{E}[\beta_{Y_i}|Z_i \in \{z - 1, z\}] = \mathrm{E}[\beta_{Y_i}] = \mathrm{ACE}$, where the second equality follows from the fact that, under NOSH, $\beta_Y \perp\!\!\!\perp (\beta_X, Z)$ (or just $\mathrm{E}[\beta_Y|Z, \beta_X] = \mathrm{E}[\beta_Y]$ under the relaxed version). This implies that the IV estimand is $\beta_{IV} = \sum_{z=2}^{K} \beta_{IV_{Z,Z-1}}\omega_z = \sum_{z=2}^{K} \mathrm{E}[\beta_{Y_i}]\omega_z = \mathrm{E}[\beta_{Y_i}]$. The case for a continuous Z is similar, where summation over Z is replaced with integration over Z: that is, $\beta_{IV} = \int_{-\infty}^{\infty} \beta_{IV_Z}\varphi_z dz$, where $\beta_{IV_Z} = \lim_{d \to 0^+} \frac{\mathrm{E}[Y_i|Z=z+d] - \mathrm{E}[Y_i|Z=z]}{\mathrm{E}[X_i|Z=z+d] - \mathrm{E}[X_i|Z=z]}$ and φ_z are weights defined such that $\int_{-\infty}^{\infty} \varphi_z dz = 1$.[3] Under NOSH, $\beta_{IV_Z} = \lim_{d \to 0^+} \mathrm{E}[\beta_{Y_i}|Z\{z, z + d\}] = \lim_{d \to 0^+} \mathrm{E}[\beta_{Y_i}] = \mathrm{E}[\beta_{Y_i}]$. Therefore, $\beta_{IV} = \int_{-\infty}^{\infty} \beta_{IV_Z}\varphi_z dz = \int_{-\infty}^{\infty} \mathrm{E}[\beta_{Y_i}]\varphi_z dz = \mathrm{E}[\beta_{Y_i}]$.

4. Simulation study

4.1. Data-generating model

We performed simulations to further demonstrate that IVs identify the ACE under NOSH, evaluating several related examples. We used the following data-generating model where Z, X and Y are continuous:

$$Z_i = \varepsilon_{Z_i} \quad (S11),$$

$$U_{k_i} \sim \text{Binom}(0.5) \text{ for } k \in \{3,4,5,6\}, \text{ independently} \quad (S12),$$

$$V_{k_i} \sim \text{Binom}(0.5) \text{ for } k \in \{3,4,5,6\}, \text{ independently} \quad (S13),$$

$$X_i \sim \gamma Z_i + \rho Z_i^3 + \sum_{k=3}^{6} \left(\delta_{X_k}^{U} U_{k_i} + \delta_{X_k}^{V} V_{k_i} + \delta_{X_k}^{U\times V} U_{k_i} V_{k_i}\right) + \sum_{k\in\{4,6\}} \left(\theta_{X_k}^{U} Z_i U_{k_i} + \theta_{X_k}^{V} Z_i V_{k_i} + \theta_{X_k}^{U\times V} Z_i U_{k_i} V_{k_i}\right) + \varepsilon_{X_i} \quad (S14),$$

$$Y_i \sim \tau X_i + \varphi X_i^2 + \sum_{k=3}^{6} \left(\delta_{Y_k}^{U} U_{k_i} + \delta_{Y_k}^{V} V_{k_i} + \delta_{Y_k}^{U\times V} U_{k_i} V_{k_i}\right) + \sum_{k\in\{5,6\}} \left(\theta_{Y_k}^{U} X_i U_{k_i} + \theta_{Y_k}^{V} X_i V_{k_i} + \theta_{Y_k}^{U\times V} X_i U_{k_i} V_{k_i}\right) + \varepsilon_{Y_i} \quad (S15),$$

where ε_{Z_i}, ε_{X_i}, ε_{Y_i} are independent, distributed as described below. k starts at 3 instead of 1 so that the classification in Table 1 also applies here.

$\{U_k\}_{k=3}^{6}$ and $\{V_k\}_{k=3}^{6}$ are two sets of independent binary confounders between the X and the Y. For all $k \in \{3,4,5,6\}$, $P(U_k = 1) = P(V_k = 1) = 0.5$. From the model, the confounders can be classified as in Table 1, as follows:

- U_3 and V_3 do not modify either the Z-X effect or the X-Y effect.
- U_4 and V_4 modify the effect of Z on X (parameters $\theta_{X_4}^{U}$, $\theta_{X_4}^{V}$ and $\theta_{X_4}^{U\times V}$ in (S14)), but not the effect of X on Y.
- U_5 and V_5 modify the effect of effect of X on Y (parameters $\theta_{Y_5}^{U}$, $\theta_{Y_5}^{V}$ and $\theta_{Y_5}^{U\times V}$ in (S15)), but not the Z-X effect.

- U_6 and V_6 modify the effect of Z on X (parameters $\theta_{X_6}^U$, $\theta_{X_6}^V$ and $\theta_{X_6}^{U\times V}$ in (S14)), and the effect of X on Y (parameters $\theta_{Y_4}^U$, $\theta_{Y_4}^V$ and $\theta_{Y_4}^{U\times V}$ in (S15)).

The data-generating model also allows the Z-X and X-Y effects to be non-linear if $\rho \neq 0$ and $\varphi \neq 0$, respectively.

4.2. Simulation scenarios

In all simulations, $\delta_{X_k}^U = \delta_{X_k}^V = \delta_{X_k}^{U\times V} = \delta_{Y_k}^U = \delta_{Y_k}^V = \delta_{Y_k}^{U\times V} = 0.5$ for all $k \in \{3,4,5,6\}$, implying that the unconditional ordinary lest squares estimate is upwardly biased for the causal effect of X on Y. Effect heterogeneity can occur on the effect of Z on X and/or on the effect of X on Y.

4.2.1. Scenario 1: NOSH holds

In scenario 1, data was generated under NOSH by setting $\theta_{X_6}^U = \theta_{X_6}^V = \theta_{X_6}^{U\times V} = \theta_{Y_6}^U = \theta_{Y_6}^V = \theta_{Y_6}^{U\times V} = \varphi = 0$. Moreover, we set $\tau = 0.5$, $\theta_{Y_5}^U = 0.5$, $\theta_{Y_5}^V = -1$ and $\theta_{Y_5}^{U\times V} = 0$, so that the effect of X on Y is linear within strata of U_5 and V_5. Since both U_5 and V_5 affect X, NEM1 and NEM2 are violated. The effect of Z on X was regulated by setting $\gamma = -0.4$, $\rho = 0.3$, $\theta_{X_4}^U = 0.5$, $\theta_{X_4}^V = -1$ and $\theta_{X_4}^{U\times V} = 0$.

4.2.2. Scenarios 2 and 3: NOSH is violated

In scenario 2, assumption 1 (but not assumption 2) is violated. This was done by setting $\theta_{X_k}^U = \theta_{X_k}^V = \theta_{X_k}^{U\times V} = \theta_{Y_k}^U = \theta_{Y_k}^V = \theta_{Y_k}^{U\times V} = \varphi = 0$ for $k \in \{4,5\}$. The causal effect of X on Y is linear within strata of U_6 and V_6, and was regulated by setting $\tau = \theta_{Y_6}^U = \theta_{Y_6}^V = \theta_{Y_6}^{U\times V} = 0.5$. The effect of Z on X was regulated by setting $\gamma = -0.4$, $\rho = 0.3$ and $\theta_{X_6}^U = \theta_{X_6}^V = \theta_{X_6}^{U\times V} = 0.5$. In scenario 3, assumption 2 (but not assumption 1) is violated. This was done by setting the parameters in the same way as in scenario 1, except that $\tau = 0$ and $\varphi = 0.2$.

4.2.3. Scenarios 4 and 5: non-normal error terms

In scenarios 1-3, ε_{Z_i}, ε_{X_i}, $\varepsilon_{Y_i} \sim N(0,1)$, independently. To corroborate the notion that NOSH does not require normal error terms, scenario 1 was re-evaluated with the difference that error terms are independently sampled from the following distributions: $\mathrm{Beta}(0.5,0.5)$ (scenario 4), $\chi^2(2) + 7I$, where $I \sim \mathrm{Bernoulli}(0.5)$ (scenario 5). In these scenarios, the error terms were converted to sample Z scores so that all have sample mean of 0 and sample variance of 1, to improve comparisons with scenario 1.

4.3. Statistical analysis

We used the two stage least squares estimator (TSLS) to estimate the ACE. For the case of a single Z, X and Y, TSLS is equivalent to the Wald estimator.[3,5]. This corresponds to assuming that NOSH holds marginally. We refer to such TSLS specification – i.e., a model with no covariates – as TSLS(1). We also considered three additional TSLS specifications, where U_6 and V_6 are incorporated in the model in different ways. To simplify the interpretation of the model, U_6 and V_6 were converted to deviations from their sample means.

- TSLS(2): including U_6 and V_6 as covariates, with no product term. That is:

$$\hat{X}_i = \widehat{\beta_0^X} + \widehat{\beta_1^X} Z_i + \widehat{\Psi_U^X} U_{6_i} + \widehat{\Psi_V^X} V_{6_i}$$

$$\hat{Y}_i = \widehat{\beta_0^Y} + \widehat{\beta_1^Y} \hat{X}_i + \widehat{\Psi_U^Y} U_{6_i} + \widehat{\Psi_V^Y} V_{6_i}$$

- TSLS(3): including U_6 and V_6, but not $U_6 \times V_6$, as covariates, with product terms with Z and X. That is:

$$\hat{X}_i = \widehat{\beta_0^X} + \widehat{\beta_1^X} Z_i + \widehat{\Psi_U^X} U_{6_i} + \widehat{\Psi_V^X} V_{6_i} + \widehat{\lambda_U^X} U_{6_i} Z_i + \widehat{\lambda_V^X} V_{6_i} Z_i$$

$$\hat{Y}_i = \widehat{\beta_0^Y} + \widehat{\beta_1^Y} \hat{X}_i + \widehat{\Psi_U^Y} U_{6_i} + \widehat{\Psi_V^Y} V_{6_i} + \widehat{\lambda_U^Y} U_{6_i} Z_i + \widehat{\lambda_V^Y} V_{6_i} Z_i$$

- TSLS(4): including U_6, V_6 and $U_6 \times V_6$ as covariates. That is:

$$\hat{X}_i = \widehat{\beta_0^X} + \widehat{\beta_1^X} Z_i + \widehat{\Psi_U^X} U_{6_i} + \widehat{\Psi_V^X} V_{6_i} + \widehat{\Psi_{U\times V}^X} V_{6_i} + \widehat{\lambda_U^X} U_{6_i} Z_i + \widehat{\lambda_V^X} V_{6_i} Z_i + \widehat{\lambda_{U\times V}^X} U_{6_i} V_{6_i} Z_i$$

$$\hat{Y}_i = \widehat{\beta_0^Y} + \widehat{\beta_1^Y} \hat{X}_i + \widehat{\Psi_U^Y} U_{6_i} + \widehat{\Psi_V^Y} V_{6_i} + \widehat{\Psi_{U\times V}^Y} V_{6_i} + \widehat{\lambda_U^Y} U_{6_i} Z_i + \widehat{\lambda_V^Y} V_{6_i} Z_i + \widehat{\lambda_{U\times V}^Y} U_{6_i} V_{6_i} Z_i$$

Assuming NOSH holds given U_6 and V_6 (as in scenario 2 of the simulation study), consistency of $\widehat{\beta_1^Y}$ as an estimate of the ACE also requires correct model specification and that covariates were measured without error. Interacting each covariate with Z and X in this way improves robustness to model specification.[6] In practice, it may be useful to compare different specifications to assess sensitivity to model misspecification.

We calculated median bias and standard error, coverage (here defined as the proportion of times that the 95% confidence intervals included the ACE) and rejection rate (here defined as the proportion of times that the 95% confidence intervals excluded the null) across 20,000 simulated datasets. Confidence intervals were calculated using Huber-White standard errors for instrumental variable analysis.[7]

5. Re-examination of selected published studies

NOSH is an untestable assumption that cannot be guaranteed by study design, and thus assessing its plausibility requires subject matter knowledge. Indeed, in some applications this assumption may be implausible. We now discuss three published IV studies to further illustrate how NOSH can be used for interpreting the Wald estimate as a consistent estimate for the ACE.

5.1. Treatment allocation in a RCT of vitamin A on mortality

Sommer et al. (1986)[8] ran a RCT investigating the effect of vitamin A supplementation (which corresponds to the treatment X) on childhood mortality (which corresponds to the treatment Y) in 450 villages in northern Sumatra (Indonesia). About half of the villages were allocated to each group at random (which corresponds to the instrument Z). 93.2% of those allocated to treatment took at least one vitamin A tablet, compared to only 1.1% of those allocated to control. The estimated effect of Z on X was thus 92.1 percent points (95% confidence interval [CI]: 91.6 to 92.7) per 100. The estimated effect of the intervention allocation on mortality was 0.29 (95% CI: 0.05 to 0.52). The Wald estimate of the effect of taking at least one pill on mortality was therefore 0.31 (95% CI: 0.05 to 0.57).

Under monotonicity, the IV estimate could be interpreted as the effect of taking at least one vitamin A tablet amongst those who would always comply to the assigned treatment. Such subgroup of the population is therefore often referred as compliers. For binary Z (in the example, $Z = 0$ and $Z = 1$ respectively denote being allocated to vitamin A or control) and X (in the example, $X = 0$ and $X = 1$ respectively denote actually taking vitamin A or not), monotonicity is defined as $X_i(Z_i = 1) \geq X_i(Z_i = 0)$ for everyone in the population, with $X_i(Z_i = z)$ defined in the same way as in section 3 in the Supplement. Individual i is a complier if $X_i(Z_i = z) = z \in \{0,1\}$.

Monotonicity is frequently proposed as an assumption that allows interpreting the IV estimate as consistent for a well-defined causal estimand (the ACE among compliers) when neither NEM nor instrument effect homogeneity hold, which are indeed often strong assumptions. However, under the assumption that the individual level effects of Z on X and of X on Y are independent, which is weaker than other IV4 assumptions known to identify the ACE (as described in section 2.3), NOSH holds (because Assumption 1 is satisfied and, because X is binary, Assumption 2 is automatically satisfied) the IV estimate can be interpreted as a consistent estimate of the ACE in the population even if other IV4 assumptions are violated. Of note, even if monotonicity is violated, the IV estimate can still be interpreted as an estimate of the ACE under NOSH (or any of the other IV4 assumptions).

In this example, NOSH requires that the participants' compliance behavior was independent of their potential outcomes. Given *a priori*, it is plausible to assume that the effects of taking vitamin A were generally not known by the study participants. However, this is an untestable assumption. A plausible situation where NOSH would be violated is if compliance were lower and the effect of vitamin A were greater among study participants from disadvantaged villages. In this case, weaker treatment effects would receive greater weight than stronger treatment effects (because the first occurs more often in subgroups of the study sample where the IV is stronger), thus leading to bias in the IV estimate if interpreted as an estimate of the ACE across the population. The magnitude of this bias will be proportional to the correlation between the individual-level effects of Z on X and of X on Y.

5.2. Using draft lottery to estimate the effect of veteran status on earnings

Angrist (1990)[9] investigated the effect of serving in the military (X) on taxable earnings (Y). He used the Vietnam draft lottery (Z) as an instrument for veteran status. Men were selected for the draft by randomly dates of birth within each year. Men whose birthday was selected were eligible for the draft; all other men could volunteer but were not drafted. After the lottery, eligible men were screened for physical and mental criteria, and some were eliminated after these screens. Men who were eligible for the draft because of the lottery were between 10 and 16 percentage points more likely to be a veteran. Eligible men had lower earnings. The Wald estimate of the effect of serving in the military on earnings was -$1920 (95%CI: -$3049 to -$792) per year in 1978 dollars.

In this example, again Assumption 2 is automatically satisfied because X is binary. NOSH would therefore require that the individual-level effects of the lottery draft on the probability of being a veteran are independent of the individual-level causal effects of veteran status on earnings. NOSH could be violated, for example, if men who were drafted, but did not become veterans, had particularly large differences in potential earnings – i.e., the causal effect is greater among those who did not serve in the military even though they were drafted (i.e., never takers and/or defiers). In case such correlation between the individual-level effects of the lottery on veteran status and the individual-level effects of serving on earnings exists, then the Wald estimate would not be consistent for the ACE.

5.3. Mendelian randomization: the effect of body mass index on coronary heart disease

As a final illustration, we discuss a Mendelian randomization study, where genetic variants robustly associated with the treatment variable are used as IVs.[4,10] Here, we discuss a study by Dale et al. (2017)[11] investigating the effects of body mass index (BMI) on coronary heart disease (CHD). They measured 97 genetic variants robustly associated with BMI in 14 prospective studies and used these

genetic variants as IVs. They combined the 97 variants into a single genetic score and found that 1 standard deviation increase in BMI increased odds of CHD by 36% (95% CI: 22% to 52%).

The genetic score is likely to have heterogenous effects on BMI across the population,[12] but it may affect everyone's BMI in the same direction. In this case, the IV estimate can be interpreted as a weighted average of the effect of increasing BMI in all those individuals whose BMI was affected by the genetic score. Under monotonicity, each individual contributes to the IV estimate proportionally to the effect of the genetic score on their BMI.[4,13] Although well-defined mathematically, this estimand is difficult to interpret from a policy-making perspective, because different unknown subgroups of the population contribute different unknown weights.

In this example, it may me implausible that NOSH holds due to the evidence supporting non-linear associations between BMI and CHD-related outcomes,[14,15] which would violate Assumption 2. Assumption 1 relates to individual-level effects of the genetic score on BMI being independent of the individual-level effects of BMI on CHD. Is this plausible? Modifiers of the effect of the genetic score on BMI may correlate with modifiers of the effect of BMI on CHD. Indeed, results from the UK Biobank indicate that the effect of the genetic score on BMI varies according to many variables, such as alcohol intake and Townsend Deprivation Index (a measure of socioeconomic position).[16] Assumption 1 would be violated if the effect of BMI on CHD varies by one of these variables.

6. References

1. Angrist JD, Imbens GW, Rubin DB. Identification of Causal Effects Using Instrumental Variables. J Am Stat Assoc. 1996;91:444–55.

2. Hartwig FP, Wang L, Smith GD, Davies NM. Homogeneity in the instrument-treatment association is not sufficient for the Wald estimand to equal the average causal effect when the exposure is continuous. Epidemiology. 2022;In press.

3. Angrist JD, Graddy K, Imbens GW. The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. Review of Economic Studies. 2000;67(3):499–527.

4. von Hinke S, Davey Smith G, Lawlor DA, Propper C, Windmeijer F. Genetic markers as instrumental variables. Journal of Health Economics. 2016;45:131–48.

5. Angrist JD, Pischke J-S. Chapter 4. Instrumental Variables in Action: Sometimes You Get What You Need. In: Mostly Harmless Econometrics. Princeton University Press; 2009. p. 113–220.

6. Imbens GW, Rubin DB. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Causal Inference: For Statistics, Social, and Biomedical Sciences an Introduction. Cambridge (England): Cambridge University Press; 2015.

7. White H. Instrumental Variables Regression with Independent Observations. Econometrica. 1982 Mar;50(2):483.

8. Sommer A, Djunaedi E, Loeden AA, Tarwotjo I, West KP, Tilden R, et al. IMPACT OF VITAMIN A SUPPLEMENTATION ON CHILDHOOD MORTALITY. A Randomised Controlled Community Trial. The Lancet. 1986;1:1169–73.

9. Angrist JD. Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. American Economic Review. 1990;80:313–36.

10. Davey Smith G, Ebrahim S. “Mendelian randomization”: Can genetic epidemiology contribute to understanding environmental determinants of disease? International Journal of Epidemiology. 2003;32:1–22.

11. Dale CE, Fatemifar G, Palmer TM, White J, Prieto-Merino D, Zabaneh D, et al. Causal Associations of Adiposity and Body Fat Distribution with Coronary Heart Disease, Stroke Subtypes, and Type 2 Diabetes Mellitus: A Mendelian Randomization Analysis. Circulation. 2017;135:2373–88.

12. Spiller W, Slichter D, Bowden J, Davey Smith G. Detecting and correcting for bias in Mendelian randomization analyses using Gene-by-Environment interactions. International Journal of Epidemiology. 2019;48(3):702–12.

13. Angrist JD, Imbens GW. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. J Am Stat Assoc. 1995;90:431–42.

14. Bhaskaran K, Dos-Santos-Silva I, Leon DA, Douglas IJ, Smeeth L. Association of BMI with overall and cause-specific mortality: a population-based cohort study of 3·6 million adults in the UK. Lancet Diabetes Endocrinol. 2018;6(12):944–53.

15. Cheng YJ, Chen ZG, Wu SH, Mei WY, Yao FJ, Zhang M, et al. Body mass index trajectories during mid to late life and risks of mortality and cardiovascular outcomes: Results from four prospective cohorts. EClinicalMedicine. 2021;33.

16. Rask-Andersen M, Karlsson T, Ek WE, Johansson Å. Gene-environment interaction study for BMI reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status. PLoS Genetics. 2017;13:e1006977.

7. Supplementary Table

Supplementary Table 1. Median bias, coverage and power of four two-stage least squares (TSLS) regression specifications as estimators of the ACE† in scenarios 1-5‡.

N	TSLS	Scenario 1 (ACE=0.250)			Scenario 2 (ACE=1.125)			Scenario 3 (ACE=0.750)			Scenario 4 (ACE=0.250)			Scenario 5 (ACE=0.250)		
		Bias	Coverage	Power	Bias	Coverage	Power	Bias	Coverage	Power	Bias	Coverage	Power	Bias	Coverage	Power
250	1	-0.001	95.4	19.4	1.262	75.3	63.4	0.407	65.8	94.3	0.001	95.6	23.1	0.003	96.7	19.6
	2	-0.002	95.1	20.0	1.261	61.0	75.1	0.407	65.2	95.4	0.001	95.5	23.3	0.003	96.4	19.7
	3	-0.010	96.3	15.9	0.144	95.2	60.9	0.380	70.9	87.8	-0.012	96.4	19.2	-0.010	97.7	14.3
	4	-0.017	97.1	13.6	0.112	99.2	35.3	0.367	74.8	80.8	-0.018	96.9	16.7	-0.016	98.2	11.7
1000	1	0.002	95.0	53.8	1.316	14.3	98.8	0.414	18.2	100.0	0.001	95.3	62.1	0.000	95.4	51.8
	2	0.002	94.7	55.7	1.314	3.4	99.2	0.414	16.3	100.0	0.001	95.3	64.4	0.001	95.5	53.8
	3	-0.001	94.8	53.1	0.143	84.1	91.0	0.407	18.1	100.0	-0.002	95.4	61.7	-0.003	95.7	50.2
	4	-0.002	94.9	51.8	0.055	99.2	49.8	0.404	19.4	100.0	-0.004	95.5	60.4	-0.004	96.0	48.1
2500	1	0.000	95.1	88.1	1.319	0.1	100.0	0.413	0.6	100.0	0.000	95.0	93.6	0.000	95.1	86.7
	2	-0.001	95.1	89.5	1.319	0.0	100.0	0.414	0.4	100.0	0.001	94.9	95.1	0.000	95.1	88.8
	3	-0.002	95.0	88.7	0.146	64.7	98.5	0.412	0.5	100.0	-0.001	95.0	94.6	-0.001	95.3	87.5
	4	-0.002	94.9	88.4	0.008	98.7	66.9	0.410	0.6	100.0	-0.001	95.1	94.5	-0.002	95.4	86.9
5000	1	0.000	95.0	99.2	1.321	0.0	100.0	0.417	0.0	100.0	0.000	95.0	99.7	0.000	95.2	98.9
	2	0.000	95.0	99.4	1.322	0.0	100.0	0.417	0.0	100.0	0.000	95.0	99.9	0.000	95.2	99.3
	3	-0.001	95.0	99.4	0.145	40.3	99.9	0.415	0.0	100.0	-0.001	95.1	99.8	0.000	95.2	99.2
	4	-0.001	95.0	99.4	-0.010	98.0	84.0	0.415	0.0	100.0	-0.001	95.0	99.9	-0.001	95.3	99.2
10000	1	0.000	94.8	100.0	1.320	0.0	100.0	0.418	0.0	100.0	0.001	94.9	100.0	0.000	95.3	100.0
	2	0.000	94.8	100.0	1.319	0.0	100.0	0.418	0.0	100.0	0.001	94.9	100.0	0.000	95.3	100.0
	3	-0.001	94.8	100.0	0.146	14.1	100.0	0.416	0.0	100.0	0.000	94.9	100.0	0.000	95.3	100.0
	4	-0.001	94.8	100.0	-0.011	97.4	96.4	0.416	0.0	100.0	0.000	95.0	100.0	0.000	95.3	100.0

†Average causal effect (ACE) of a unit increase in X on Y.

‡1: NOSH holds. 2: Assumption 1 violated. 3: Assumption 2 violated. 4: NOSH holds and error terms are drawn from a beta distribution. 5: NOSH holds and error terms are drawn from a mixed chi-squared distribution.