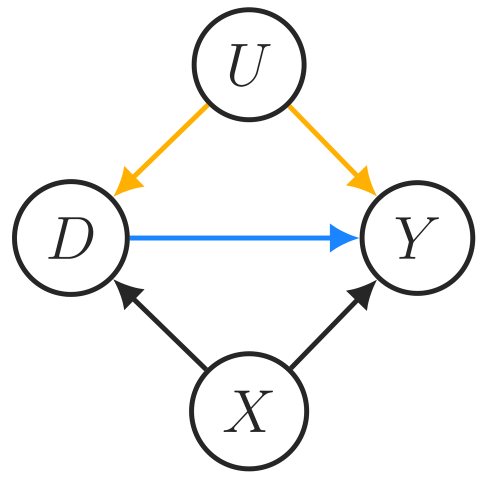
Sensitivity Analysis with the R^2-Calculus

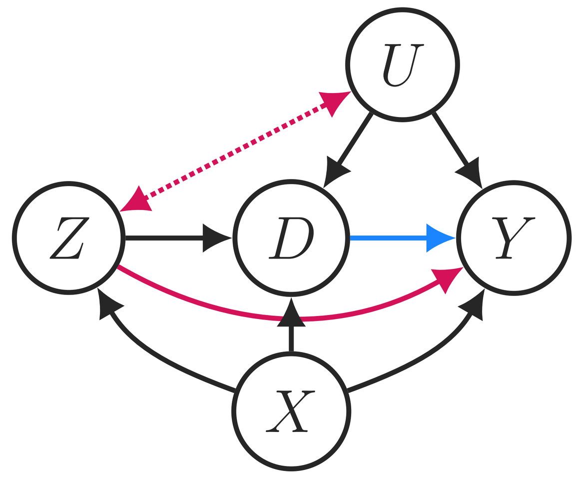
Tobias Freidling[1,2], Qingyuan Zhao[1,2]

[1]DPMMS, University of Cambridge
[2]Cantab Capital Institute for the Mathematics of Information

Untestable Assumptions

To estimate the true effect (blue arrow) of a variable D on an outcome Y, additional, unverifiable assumptions are needed.
For instance, linear regression requires that an unmeasured confounder U does not effect D and Y (yellow arrows) simultaneously. In an instrumental variable setting, it is required that the instrument Z influences Y only through D and that it is independent of U, i.e. absence of the red arrows.
Sensitivity analysis allows practitioners to explore unmeasured confounding and its effect on the estimate and confidence interval.





R^2-Calculus

In a linear regression $Y = X\beta + \varepsilon$, the coefficient of determination, i.e. $R^2_{Y\sim X}$, is the proportion of variance in Y that is explained by the model.

Definitions

Let $A \in \mathbb{R}$, $B \in \mathbb{R}^d$, $C \in \mathbb{R}^k$ and $D \in R^l$ be random vectors

- R^2-value: $R^2_{A\sim B} = 1 - \frac{\mathrm{var}(A - B\beta)}{\mathrm{var}(A)}$, where β is the regression coefficient.
- Partial R^2-value: $R^2_{A\sim B|C} = \frac{R^2_{A\sim B+C} - R^2_{A\sim C}}{1 - R^2_{A\sim C}}$.
- R-value: $R_{A\sim B|C} = \mathrm{corr}(A^{\perp C}, B^{\perp C})$, for $B \in \mathbb{R}$.
- (Partial) f-value: $f^2_{A\sim B|C} = \frac{R^2_{A\sim B|C}}{1 - R^2_{A\sim B|C}}$; $\quad f_{A\sim B|C} = \frac{R_{A\sim B|C}}{\sqrt{1 - R^2_{A\sim B|C}}}$, for $B \in \mathbb{R}$.

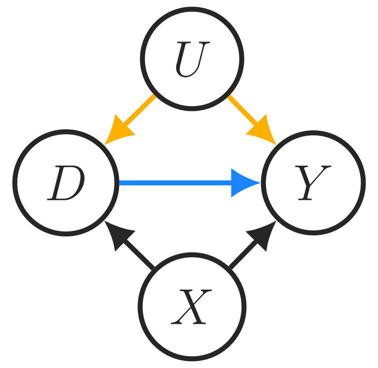
Calculation Rules

- If $B \perp\!\!\!\perp C$, then $R^2_{A\sim B+C} = R^2_{A\sim B} + R^2_{A\sim C}$.
- $\frac{\mathrm{var}(A^{\perp B,C})}{\mathrm{var}(A^{\perp C})} = 1 - R^2_{A\sim B|C}$ and thus $1 - R^2_{A\sim B+C|D} = (1 - R^2_{A\sim B|D})(1 - R^2_{A\sim C|B,D})$
- $R_{A\sim B|C,D} = \frac{R_{A\sim B|C} - R_{A\sim D|C} R_{B\sim D|C}}{\sqrt{1 - R^2_{A\sim D|C}}\sqrt{1 - R^2_{B\sim D|C}}}$, for $A, B, D \in \mathbb{R}$.
- If $A, B \in \mathbb{R}$ and $A \perp\!\!\!\perp (D, C)$, then $R_{A\sim B|C,D} = \frac{R_{A\sim B|D}}{\sqrt{1 - R^2_{B\sim C|D}}}$.

Linear Regression

A linear regression model[1] with one-dimensional outcome Y, variable of interest D and unmeasured confounder U and multi-dimensional covariates X is given by

$$Y = D\beta + U\gamma + \lambda^T X + \varepsilon.$$



The bias in the β-estimate when excluding U can be expressed with the R^2-calculus as

$$\mathrm{bias} = R_{Y\sim U|D,X}\, f_{D\sim U|X} \frac{\mathrm{sd}(Y^{\perp D,X})}{\mathrm{sd}(D^{\perp X})}.$$

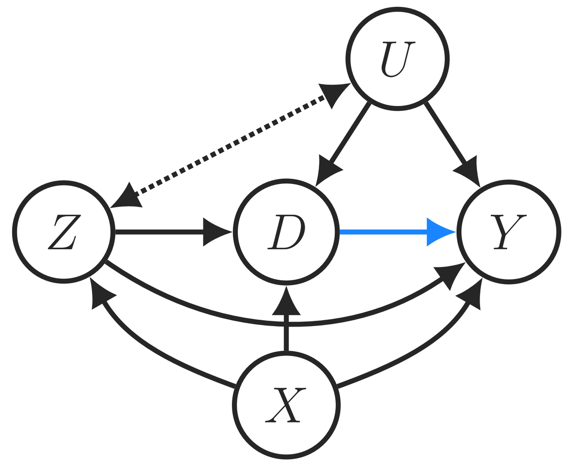
This allows a practitioner to find a range for the bias by reasoning about the two unobservable quantities $R_{Y\sim U|D,X}$ and $f_{D\sim U|X}$. For instance, one can specify the inequality $R^2_{D\sim U} \leq 0.5 R^2_{D\sim X_j}$ which yields a bound on $f_{D\sim U|X}$ using the R^2-calculus.
If the ranges $R_{Y\sim U|D,X} =: R \in [b_R^-, b_R^+]$ and $f_{D\sim U|X} =: f \in [b_f^-, b_f^+]$ are specified, the maximum/minimum bias is

$$\max_{R\in[b_R^-, b_R^+],\, f\in[b_f^-, b_f^+]} \Big/ \min_{R\in[b_R^-, b_R^+],\, f\in[b_f^-, b_f^+]} R_{Y\sim U|D,X}\, f_{D\sim U|X} \frac{\mathrm{sd}(Y^{\perp D,X})}{\mathrm{sd}(D^{\perp X})}.$$

K-class Estimator

The κ-class estimate for a linear system depicted in the graph on the right is given by

$$\beta_\kappa = \frac{\mathrm{cov}(D^{\perp X}, Y^{\perp X}) - \kappa\, \mathrm{cov}(D^{\perp Z,X}, Y^{\perp Z,X})}{\mathrm{var}(D^{\perp X}) - \kappa\, \mathrm{var}(D^{\perp Z,X})}$$



and interpolates between the IV estimate, that is $\kappa = 1$, and the estimates from the linear regressions $Y \sim D + X$, i.e. $\kappa = 0$, and $Y \sim D + X + Z$, i.e. $\kappa \to -\infty$, respectively.
Omitting U[2] in the estimation leads to the bias

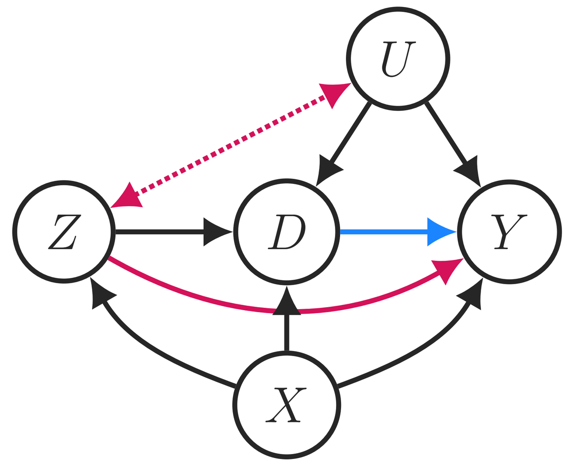
$$\mathrm{bias} = \left[\frac{f_{Y\sim Z|D,X}\, R_{D\sim Z|X}}{1 - \kappa\,(1 - R^2_{D\sim Z|X})} + R_{Y\sim U|D,Z,X}\, f_{D\sim U|Z,X}\right] \frac{\mathrm{sd}(Y^{\perp D,Z,X})}{\mathrm{sd}(D^{\perp Z,X})}.$$

Remarkably, in order to bound the bias it suffices to specify bounds for just two quantities: $R_{Y\sim U|D,Z,X}$ and $f_{D\sim U|Z,X}$, which parametrise the linear regression assumptions.
This result can be extended to a multi-dimensional Z with independent components, see also [3].

Instrumental Variables

In a linear Instrumental Variable (IV) model, the estimate

$$\beta_{\mathrm{IV}} = \frac{\mathrm{cov}(Z, Y)}{\mathrm{cov}(Z, D)}$$



is unbiased if the instrument $Z \in \mathbb{R}$ influences Y only through D and $Z \perp\!\!\!\perp U$, even in the presence of an unmeasured confounder.
When these assumptions are violated, the bias is

$$\mathrm{bias} = \left[\frac{R_{Y\sim U|D,Z,X}\, f_{U\sim Z|X}}{f_{D\sim Z|X}\sqrt{1 - R^2_{D\sim U|Z,X}}} + \frac{R_{Y\sim Z|D,U,X}\sqrt{1 - R^2_{Y\sim U|D,X}}}{R_{D\sim Z|X}\sqrt{1 - R^2_{Z\sim U|D,X}}\sqrt{1 - R^2_{Y\sim Z|D,X}}}\right] \frac{\mathrm{sd}(Y^{\perp D,Z,X})}{\mathrm{sd}(D^{\perp Z,X})}.$$

To perform sensitivity analysis, we require a practitioner to specify bounds for $R_{U\sim Z|X}$ and $R_{Y\sim Z|D,U,X}$ which parametrise the IV assumptions and one additional parameter, e.g. $R_{Y\sim U|D,Z,X}$. Hence, the following three equations implicitly constrain the bias

$$R_{Y\sim U|D,Z,X} = \frac{R_{Y\sim U|D,X} - R_{Y\sim Z|D,X} R_{Z\sim U|D,X}}{\sqrt{1 - R^2_{Y\sim Z|D,X}}\sqrt{1 - R^2_{Z\sim U|D,X}}}, \quad R_{Y\sim Z|D,U,X} = \frac{R_{Y\sim Z|D,X} - R_{Y\sim U|D,X} R_{Z\sim U|D,X}}{\sqrt{1 - R^2_{Y\sim U|D,X}}\sqrt{1 - R^2_{Z\sim U|D,X}}},$$

$$f_{Z\sim U|D,X} = f_{Z\sim U|X}\sqrt{\frac{1 - R^2_{D\sim Z|X}}{1 - R^2_{D\sim U|Z,X}}} - R_{Z\sim D|X}\, f_{D\sim U|Z,X}.$$

Discussion and Outlook

The R^2-calculus is a comprehensive framework to analyse confounding in linear models. Importantly, it allows practitioners to think about the sensitivity parameters in an intuitive fashion.
Future work includes sensitivity analysis for confidence intervals as well as generalising the setting to multiple confounders[4]. To analyse more involved systems, implementing the R^2-calculus in a computer algebra system seems to be necessary.

References

[1] Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, 2020.

[2] Judea Pearl. On a class of bias-amplifying variables that endanger effect estimates. *arXiv*, (1203.3503), 2012.

[3] Dylan S Small. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102(479):1049–1058, 2007.

[4] Carrie A. Hosman, Ben B. Hansen, and Paul W. Holland. The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *The Annals of Applied Statistics*, 4(2):849 – 870, 2010.

E-mail: taf40@cam.ac.uk

CCIMI - Industry Engagement

Website: tobias-freidling.onrender.com