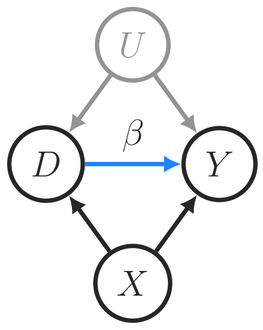
Sensitivity Analysis with the R^2-Calculus

Tobias Freidling, Qingyuan Zhao

'All models are wrong, but some are useful.'

Statistical models help us understand data but their correctness relies on unverifiable assumptions.

Example: To estimate the effect β of education D on wages Y, we gather data on these quantities and potential confounders X, such as age, parental education etc. Yet, there will be variables U, e.g. ability, that are not part of the data and hence the model. The estimate of β is only correct if U was included.



Sensitivity Analysis allows researchers to assess how robust the conclusions of their models are. This work concentrates on linear regression and instrumental variables.

R^2-Calculus

The R^2-value $R^2_{Y\sim X}$ is the proportion of variance in Y that is explained by X. It takes values in $[0,1]$ and is easy to interpret for researchers.

Calculation Rules

Let $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$ and $W \in \mathbb{R}^{n\times r}$ be n i.i.d. samples.

- R^2-value and partial R^2-value:

$$R^2_{Y\sim X} := 1 - \frac{\operatorname{var}(Y - X\hat{\beta}_X)}{\operatorname{var}(Y)}, \qquad R^2_{Y\sim X|W} := \frac{R^2_{Y\sim X+W} - R^2_{Y\sim W}}{1 - R^2_{Y\sim W}}.$$

- R-value: $R_{Y\sim X} := \operatorname{corr}(Y, X)$, for $X \in \mathbb{R}^n$.
- f^2-value and f-value:

$$f^2_{Y\sim X} := \frac{R^2_{Y\sim X}}{1 - R^2_{Y\sim X}}, \qquad f_{Y\sim X} := \frac{R_{Y\sim X}}{\sqrt{1 - R^2_{Y\sim X}}}.$$

- If $X \perp\!\!\!\perp Z$, then $R^2_{Y\sim X+Z} = R^2_{Y\sim X} + R^2_{Y\sim Z}$.
- $\operatorname{var}(Y^{\perp X})/\operatorname{var}(Y) = 1 - R^2_{Y\sim X}$.
- $1 - R^2_{Y\sim X+W} = (1 - R^2_{Y\sim X|W})(1 - R^2_{Y\sim W})$.
- For $X, W \in \mathbb{R}^n$,

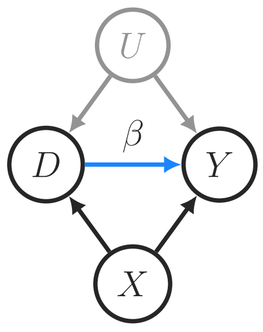
$$R_{Y\sim X|W} = \frac{R_{Y\sim X} - R_{Y\sim W}R_{X\sim W}}{\sqrt{1 - R^2_{Y\sim W}}\sqrt{1 - R^2_{X\sim W}}}.$$

- If $X \in \mathbb{R}^n$ and $Y \perp\!\!\!\perp W$, then $R_{Y\sim X|W} = R_{Y\sim X}/\sqrt{1 - R^2_{X\sim W}}$.

Linear Regression

A linear regression model[1],[2] with one-dimensional outcome Y, variable of interest D and unmeasured confounder U and multi-dimensional covariates X is given by

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



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

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

If a researcher has believes about the unmeasured confounder U and specifies respective constraints (C), we find the maximal bias by solving

$$\max_{R_{Y\sim U|X,D},\, R_{D\sim U|X}} R_{Y\sim U|X,D}\, f_{D\sim U|X}\, \frac{\operatorname{sd}(Y^{\perp X,D})}{\operatorname{sd}(D^{\perp X})}, \qquad \text{subject to } (C).$$

The same is possible for the minimal bias and confidence intervals. There are many different options to specify (C). For instance, the inequality $R^2_{D\sim U} \leq 0.5 R^2_{D\sim X_j}$ encapsulates the belief that U can explain at most half as much variance in D as X_j can.

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] 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.

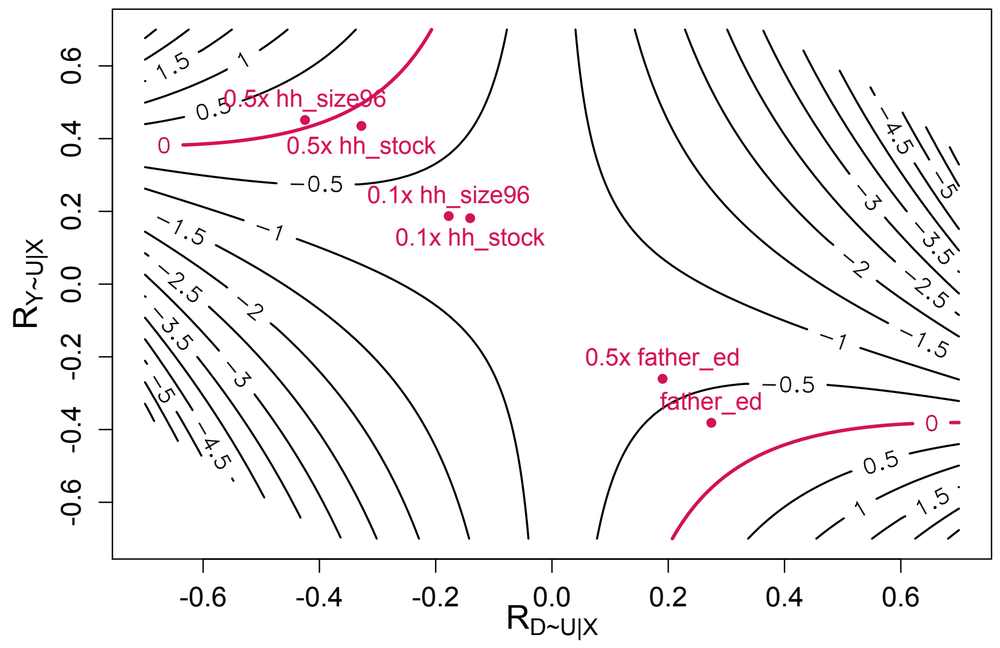
[3] Christopher Blattman and Jeannie Annan. The Consequences of Child Soldiering. *Review of Economics and Statistics*, 92(4):882–898, 2010.

Case Study: Child Soldiers in Uganda

- Background From the late 1980's to the mid-2000's, the Lord's Resistance Army (LRA), led by Joseph Kony, abducted 60,000 to 80,000 children in Northern Uganda, forced them to become chield soldiers and commit or witness atrocities. Most could escape, but how are their lives affected?
- Data The first phase of the Survey of War Affected Youth[3] was conducted in 2006 and contains data of 741 males concerning their living conditions in 1996 and 2006. 62% of them were abducted.
- Effect on Education According to a linear regression model, abduction D accounts for the loss of 0.82 years of education Y (95% confidence interval: [-1.25, -0.39]). How robust is this result in the presence of an unmeasured confounder U, e.g. mental ability or cunningness?
- Sensitivity Analysis Suppose we believe
  - $R^2_{D\sim U|X_{-j}} \leq 0.1\, R^2_{D\sim X_j|X_{-j}}$: U can explain at most 10% as much variance in D as the house hold size X_j can.
  - $R^2_{Y\sim U|X_{-l},D} \leq 0.25\, R^2_{Y\sim X_l|X_{-l},D}$: U can explain at most 25% as much variance in Y as the education of the father X_l can.

  The corresponding sensitivity interval is [-2.11, 0.46]. Hence, if the assumptions on the unmeasured confounder are true, we cannot conclude that the effect of abduction on education is significant.

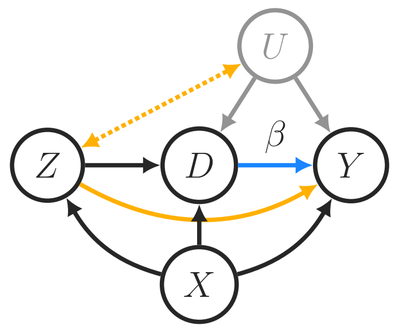
Sensitivity contours show how strong U must be to push the upper end of the sensitivity interval beyond 0 and compare it to other variables.



Instrumental Variables

In a linear Instrumental Variable (IV) model, an instrument Z is used to estimate the effect of D on Y:

$$\hat{\beta}_{\text{IV}} = \frac{\operatorname{cov}(Z^{\perp X}, Y^{\perp X})}{\operatorname{cov}(Z^{\perp X}, D^{\perp X})}.$$



Under the assumption that Z is valid, i.e. absence of the yellow arrows, $\hat{\beta}_{\text{IV}}$ is unbiased, even under unmeasured confounding.
If this assumption is violated, the ensuing bias is

$$\text{bias} = \left[\frac{f_{Y\sim Z|X,D}}{R_{D\sim Z|X}} + R_{Y\sim U|X,Z,D}\, f_{D\sim U|X,Z}\right] \frac{\operatorname{sd}(Y^{\perp X,Z,D})}{\operatorname{sd}(D^{\perp X,Z})}.$$

Similarly to linear regression, the range of the bias can be computed by specifying constraints (C). In the context of IV models, it is however unusual to reason about $R_{Y\sim U|X,Z,D}$ or $R_{D\sim U|X,Z}$. Via the rules of the R^2-calculus, we find the equations

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

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

which contain the more interpretable quantities $R_{Z\sim U|X}$ and $R_{Y\sim Z|X,U,D}$. Thus, adding the equations to (C), researchers have more intuitive options to specify bounds on sensitivity parameters.

E-mail: taf40@cam.ac.uk

tobias-freidling.onrender.com