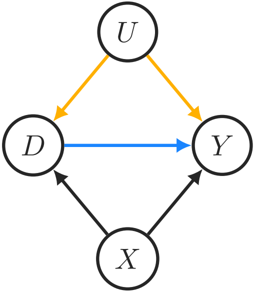
Sensitivity Analysis with the R^2-Calculus

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

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

Untestable Assumptions

Regression



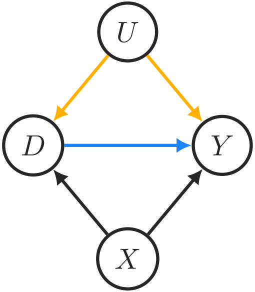
There is no unmeasured confounder U, i.e. U cannot effect D and Y (yellow arrows) simultaneously.

Instrumental Variables

The instrument Z influences Y only through D and it is independent of U, that is absence of the red arrows.

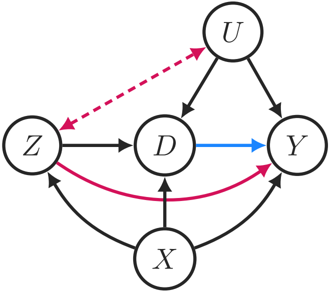
Untestable Assumptions

Regression



There is no unmeasured confounder U, i.e. U cannot effect D and Y (yellow arrows) simultaneously.

Instrumental Variables



The instrument Z influences Y only through D and it is independent of U, that is absence of the red arrows.

R^2-Calculus

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

R^2-Calculus

Let $Y \in \mathbb{R}$, $X \in \mathbb{R}^d$, $Z \in \mathbb{R}^k$ and $W \in \mathbb{R}^l$ be random vectors

- $R^2_{Y\sim X} = 1 - \frac{\mathrm{var}(Y - X\beta)}{\mathrm{var}(Y)}$, where β is the regression coefficient
- $R^2_{Y\sim X|Z} = \frac{R^2_{Y\sim X+Z} - R^2_{Y\sim Z}}{1 - R^2_{Y\sim Z}}$
- $\frac{\mathrm{var}(Y^{\perp X,Z})}{\mathrm{var}(Y^{\perp Z})} = 1 - R^2_{Y\sim X|Z}$
- $R_{Y\sim X|Z} = \mathrm{corr}(Y^{\perp Z}, X^{\perp Z})$, for $X \in \mathbb{R}$
- $R_{Y\sim X|Z,W} = \frac{R_{Y\sim X|Z} - R_{Y\sim W|Z} R_{X\sim W|Z}}{\sqrt{1 - R^2_{Y\sim W|Z}}\sqrt{1 - R^2_{X\sim W|Z}}}$, for $X, W \in \mathbb{R}$.
- $f^2_{Y\sim X|Z} = \frac{R^2_{Y\sim X|Z}}{1 - R^2_{Y\sim X|Z}}$; $\quad f_{Y\sim X|Z} = \frac{R_{Y\sim X|Z}}{\sqrt{1 - R^2_{Y\sim X|Z}}}$, for $X \in \mathbb{R}$

R^2-Calculus

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

R^2-Calculus

Let $Y \in \mathbb{R}$, $X \in \mathbb{R}^d$, $Z \in \mathbb{R}^k$ and $W \in \mathbb{R}^l$ be random vectors

- $R^2_{Y\sim X} = 1 - \frac{\mathrm{var}(Y - X\beta)}{\mathrm{var}(Y)}$, where β is the regression coefficient
- $R^2_{Y\sim X|Z} = \frac{R^2_{Y\sim X+Z} - R^2_{Y\sim Z}}{1 - R^2_{Y\sim Z}}$
- $\frac{\mathrm{var}(Y^{\perp X,Z})}{\mathrm{var}(Y^{\perp Z})} = 1 - R^2_{Y\sim X|Z}$
- $R_{Y\sim X|Z} = \mathrm{corr}(Y^{\perp Z}, X^{\perp Z})$, for $X \in \mathbb{R}$
- $R_{Y\sim X|Z,W} = \frac{R_{Y\sim X|Z} - R_{Y\sim W|Z} R_{X\sim W|Z}}{\sqrt{1 - R^2_{Y\sim W|Z}}\sqrt{1 - R^2_{X\sim W|Z}}}$, for $X, W \in \mathbb{R}$.
- $f^2_{Y\sim X|Z} = \frac{R^2_{Y\sim X|Z}}{1 - R^2_{Y\sim X|Z}}$; $\quad f_{Y\sim X|Z} = \frac{R_{Y\sim X|Z}}{\sqrt{1 - R^2_{Y\sim X|Z}}}$, for $X \in \mathbb{R}$

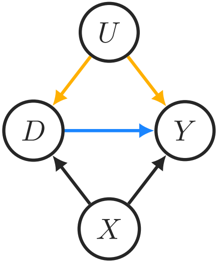
Sensitivity Analysis - Linear Regression

Linear regression model:

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

Bias in the β-estimate when excluding U:

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



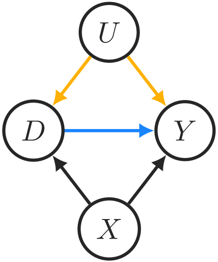
We can find a range for the bias by reasoning about $R_{Y\sim U|D,X}$ and $f_{D\sim U|X}$. For instance, if a researcher believes $R^2_{D\sim U} \leq 0.5\, R^2_{D\sim X}$, we apply the rules of the R^2-calculus and find the bound

$$|f_{D\sim U|X}| \leq \sqrt{\frac{0.5 f^2_{D\sim X}}{1 - 0.5 f^2_{D\sim X}}}.$$

Sensitivity Analysis - Linear Regression

Linear regression model:

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



Bias in the β-estimate when excluding U:

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

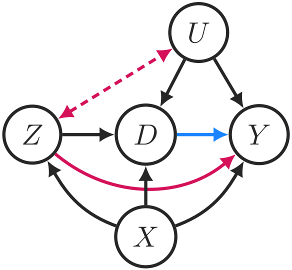
We can find a range for the bias by reasoning about $R_{Y\sim U|D,X}$ and $f_{D\sim U|X}$. For instance, if a researcher believes $R^2_{D\sim U} \leq 0.5\, R^2_{D\sim X}$, we apply the rules of the R^2-calculus and find the bound

$$|f_{D\sim U|X}| \leq \sqrt{\frac{0.5 f^2_{D\sim X}}{1 - 0.5 f^2_{D\sim X}}}.$$

Sensitivity Analysis - Instrumental Variables

Linear Instrumental Variables model:

$$D = Z\theta + U\gamma + \lambda^T X + \varepsilon_D$$
$$Y = D\beta + U\tilde{\gamma} + \tilde{\lambda}^T X + Z\tilde{\theta} + \varepsilon_Y$$



The estimate

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

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.

Sensitivity Analysis - Instrumental Variables

Bias under violation of the assumptions (dropping conditioning on X):

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

The values $R_{U\sim Z}$ and $R_{Y\sim Z|D,U}$ correspond to the IV assumptions. For sensitivity analysis, we need a bound on one additional parameter, for example $R_{Y\sim U|D,Z}$.

The unknown terms in the bias are implicitely specified by

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

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

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

Sensitivity Analysis - Instrumental Variables

Bias under violation of the assumptions (dropping conditioning on X):

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

The values $R_{U\sim Z}$ and $R_{Y\sim Z|D,U}$ correspond to the IV assumptions. For sensitivity analysis, we need a bound on one additional parameter, for example $R_{Y\sim U|D,Z}$.
The unknown terms in the bias are implicitely specified by

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

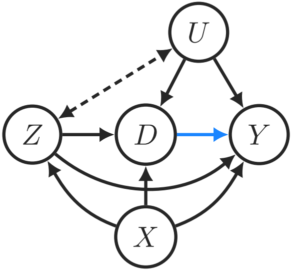
$$R_{Y\sim Z|D,U} = \frac{R_{Y\sim Z|D} - R_{Y\sim U|D}\, R_{Z\sim U|D}}{\sqrt{1-R^2_{Y\sim U|D}}\sqrt{1-R^2_{Z\sim U|D}}}$$

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

Sensitivity Analysis - K-class estimation

K-class estimate for a linear IV model:

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



Interpolation:

- $\kappa = 1$: IV estimate
- $\kappa = 0$: regression estimate of $Y \sim D + X$
- $\kappa \to -\infty$: regression estimate of $Y \sim D + X + Z$

Bias under violation of IV and regression assumptions:

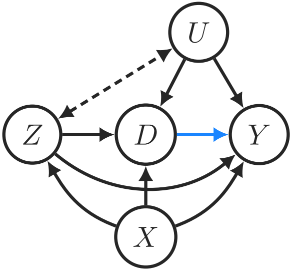
$$\text{bias} = \left[\frac{f_{Y \sim Z|D,X}\, R_{D \sim Z|X}}{1 - \kappa\,(1 - R^2_{D \sim Z|X})} + R_{Y \sim U|D,Z,X}\, f_{D \sim U|Z,X} \right] \frac{\mathrm{sd}(Y^{\perp D,Z,X})}{\mathrm{sd}(D^{\perp Z,X})}$$

It suffices to specify bounds for two quantities: $R_{Y \sim U|D,Z,X}$ and $f_{D \sim U|Z,X}$. This extends to multiple independent instruments.

Sensitivity Analysis - K-class estimation

K-class estimate for a linear IV model:

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



Interpolation:

- $\kappa = 1$: IV estimate
- $\kappa = 0$: regression estimate of $Y \sim D + X$
- $\kappa \to -\infty$: regression estimate of $Y \sim D + X + Z$

Bias under violation of IV and regression assumptions:

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

It suffices to specify bounds for two quantities: $R_{Y\sim U|D,Z,X}$ and $f_{D\sim U|Z,X}$. This extends to multiple independent instruments.

Outlook

Short-term:

- Multiple unmeasured confounders: An upper bound for the bias in linear regression is already known.
- "Combination" of the bounds for different sensitivity parameters: Do we want to allow simultaneous worst-case violations for multiple parameters?
- Application to real-world data, e.g. in econometrics

Long-term:

- Computer algebra system for the R^2-calculus
- Properties of R^2-calculus, e.g. what is the minimum number of sensitivity parameters for a given model?
- Generalisation of R^2-values

Outlook

Short-term:

- Multiple unmeasured confounders: An upper bound for the bias in linear regression is already known.
- "Combination" of the bounds for different sensitivity parameters: Do we want to allow simultaneous worst-case violations for multiple parameters?
- Application to real-world data, e.g. in econometrics

Long-term:

- Computer algebra system for the R^2-calculus
- Properties of R^2-calculus, e.g. what is the minimum number of sensitivity parameters for a given model?
- Generalisation of R^2-values

References

Cinelli, Carlos and Chad Hazlett (2020). "Making sense of sensitivity: extending omitted variable bias". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.1, pp. 39–67.

Hosman, Carrie A., Ben B. Hansen, and Paul W. Holland (2010). "The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder". In: *The Annals of Applied Statistics* 4.2, pp. 849 –870.

Pearl, Judea (2012). "On a Class of Bias-Amplifying Variables that Endanger Effect Estimates". In: *arXiv* 1203.3503.

Small, Dylan S (2007). "Sensitivity Analysis for Instrumental Variables Regression With Overidentifying Restrictions". In: *Journal of the American Statistical Association* 102.479, pp. 1049–1058.