効果検証型の実証研究におけるリサーチデザイン

講師：遠山 祐太

最終更新：2023-05-25

はじめに

テーマ：効果検証のための計量経済学

- ミクロデータを用いた実証研究の多くは何かしらの**効果**を推定することを目的としている。
- そのために、線形回帰モデルを考え、興味ある変数の係数を推定する.

$$y_i = \alpha_0 + \alpha D_i + \beta x_i + \epsilon_i$$

  - D_i: 興味ある変数、 x_i :その他の変数
  - 知りたいもの： α_1

- しかし「OLSで推定した係数」が「ある変数の効果」の良い推定値であるかは議論が必要！
- 本資料の目的：
  - **計量経済学の基礎的な知識を前提とした**効果測定のための実証分析のガイダンス
  - 統計学・計量経済学における**因果推論**の(独自の味付けをした)入門

資料の流れ

1. 「相関」と「因果」
2. リサーチデザイン、識別戦略とは？
3. 具体的なアプローチ・手法
4. 研究遂行に当たってのアドバイス

参考資料：

- 技術面も含めた詳細については、秋学期「応用計量経済学-経済学における因果推論-」にて。
- 前に配布した「ミクロ実証研究のはじめ方」はより俯瞰的なガイダンス資料になっている。

「相関」と「因果」

知りたいこと：因果関係

- **因果関係**：(他の要因を固定したときに)XがYを引き起こす

- 経済学・政治学における多くの問いは因果関係に関するもの！
  - 教育年数が一年増えると賃金はどれだけ上昇するか？
  - 商品の売れ行きにネット広告はどれだけ影響するか？
  - 企業の合併は商品価格を上昇させるか？
  - 民主主義は経済成長を引き起こすか？
  - 投票率の向上は大統領選で民主党に益するか？

データ分析の第一歩としての記述分析

- 分析はデータの**記述**から始まる。
  - 大卒者はその他に比べて時給が98%高い。
  - 現在の所得不平等は30年前に比べて大きい。
  - 医療制度改革後の医療費の伸びは緩やか。
  - 現在の航空券価格は合併以前より高い。
- 次なる問：発見したデータにおけるパターンを**どのように解釈するか？**
- **データで発見した相関を因果として解釈できるのか？**

キーポイント：相関は因果を意味するか？

- データを見たときにXとYが共に変化しているとしよう（**相関**）

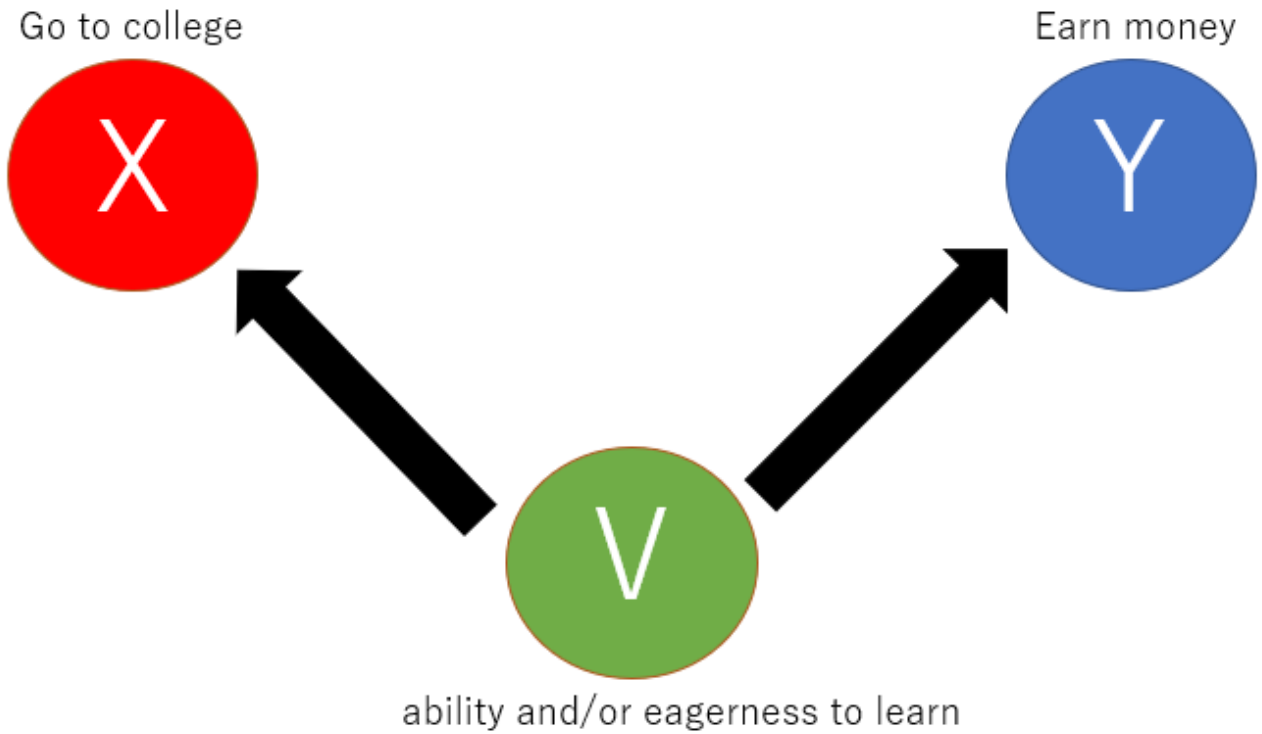
- 例：
  - １：警察官の多い都市は犯罪も多い。（正の相関）
  - ２：大学に進学した人の方が10%収入が多い。

- 問：
  - これは本当に「XはYを引き起こす」ことを意味するだろうか？
  - 相関の大きさは、因果効果の大きさとして解釈できるだろうか？

三つの可能性

1. 因果(知りたいもの)
2. 逆因果
3. 第三の要因：**交絡因子（confounder）**

三つの可能性を図で


Go to college
X
Earn money
Y
V
ability and/or eagerness to learn

線形回帰分析＝相関関係

- 線形回帰モデルの推定量も、データにおける相関関係（共分散）
- 例えば $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$における、OLS推定量は

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

  これは x と y の(標本)共分散を(標本)分散で割ったもの

- もしモデル $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ が y と x の関係を捉えるならば、 β_1 は因果効果。
- では、OLS推定量 $\hat{\beta}_1$ が 真のパラメタ β_1 を正しく推定するための条件とは？

欠落変数バイアス (Omitted Variable Bias, OVB)

- 以下の賃金回帰モデルを考える (真のモデル)

$$\begin{aligned} \log W_i &= \beta_0 + \beta_1 S_i + \beta_2 A_i + u_i \\ E[u_i | S_i, A_i] &= 0 \end{aligned}$$

  W_i は賃金、 S_i は教育年数、 A_i は能力

- β_1: **他の要因を固定した上での** 教育年数が賃金に与える効果
- 問題点：能力を直接観察できるとは限らない。

能力 A_i を除いた回帰モデル

- 以下の回帰モデルを考える。

$$\log W_i = \alpha_0 + \alpha_1 S_i + v_i$$

  - $v_i = \beta_2 A_i + u_i$, 従って、 S_i and v_i はおそらく相関している。

- このモデルで推定した $\hat{\alpha}_1$ は β_1 についてInconsistentとなる。

$$\hat{\alpha}_1 \xrightarrow{p} \beta_1 + \beta_2 \frac{Cov(S_i, A_i)}{Var(S_i)}$$

欠落変数バイアスの公式

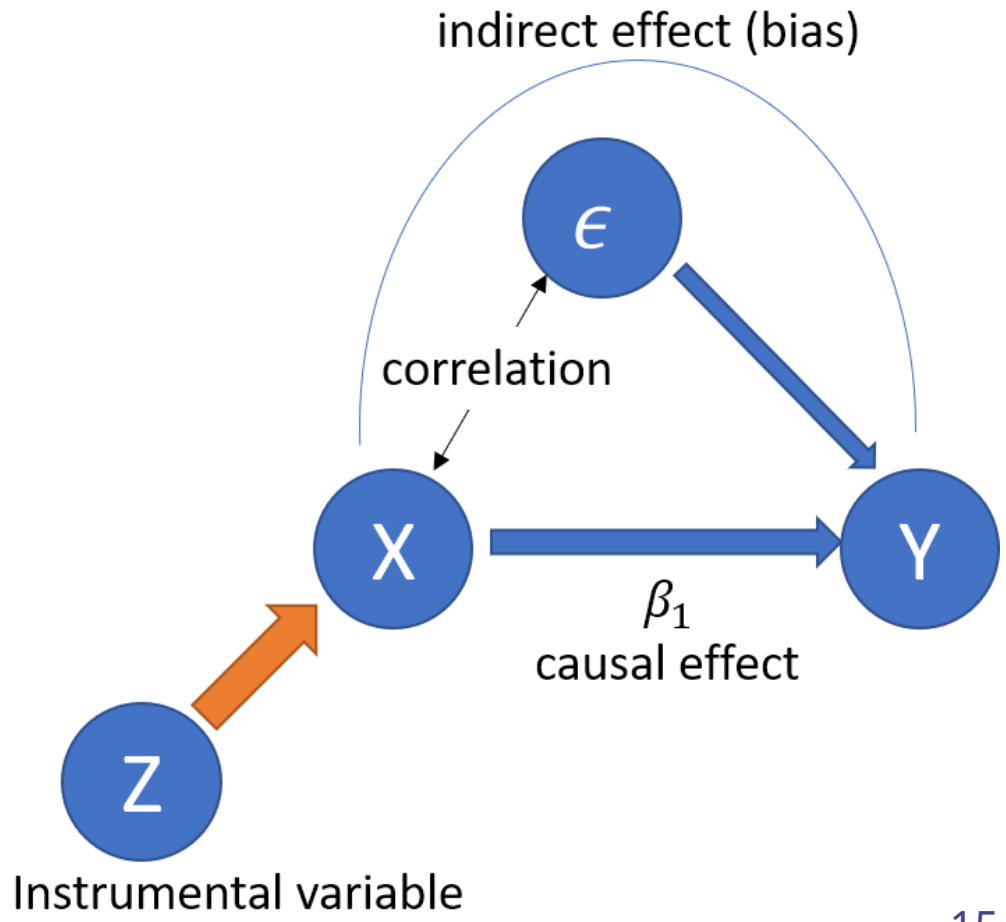
- 欠落変数バイアスは以下の２つに依存する。
  1. 欠落変数 (ここでは A_i) が 左辺に与える影響: β_2
  2. 欠落変数 と 興味ある説明変数の関係

- まとめ
  - x_1: 興味ある変数, x_2 欠落変数
  - β_2 は x_2 の係数.

	$Cov(x_1, x_2) > 0$	$Cov(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

欠落変数バイアスの直観

- OLS は x と y の相関を捉える。
- もし x と ϵ に相関がないならば、OLSは因果効果 β_1 を捉える。
- 相関がある場合、OLS は因果効果 β_1 と間接効果(バイアス)を捉える。



因果効果と反実仮想

- 因果効果を知るためには、究極的には**「反実仮想」**を知る必要がある。
- 反実仮想：「効果を知りたい変数が別の値であった時、何が起きたか？」
- ポイント：個人レベルでは、反実仮想は決して知りうることができない！！

例: 大学進学

- Rubinの**潜在アウトカムモデル**に基づく例
- ノーテーション
  - D_i 大学に進学したか否かダミー変数
  - Y_{1i}: 大学進学した場合の潜在アウトカム
  - Y_{0i}: 大学進学しなかった場合の潜在アウトカム
  - Y_i: 実際の観察される所得

	Y_{1i}	Y_{0i}	D_i	Y_i
Adam	80000 USD	50000 USD	1	80000 USD
Bob	60000 USD	60000 USD	0	60000 USD
Cindy	90000 USD	60000 USD	1	90000 USD
Debora	80000 USD	70000 USD	0	70000 USD

反実仮想の推定

- 個人レベルの反実仮想を原理的に知ることはできない。
- 次善策：反実仮想を「推定」しよう。
- 一つの案：大学進学した人の「もし大学に進学しなかった場合」の所得を、「大学に進学しなかった人たちの所得」で置き換えてはどうか？
- これは「反実仮想」の推定として良い方法か？

次の節に行く前に：用語＆概念の整理

- 因果推論においては、二値の処置（トリートメント）における効果の推定に着目
  - コントロール・グループ（対照群）：処置を受けていない人
  - トリートメント・グループ（処置群）：処置を受けた人

- 線形回帰分析では、効果を知りたい興味ある変数が必ずしも二値ではない。
  - 例：教育年数

- いずれのケースでも、因果効果の推定に関する議論の本質は同じ。以下では区別せずに話を進める。

リサーチデザイン、識別戦略

リサーチデザイン(Research Design)

- **リサーチデザイン**：観察データから因果効果をバイアスなく推定するための分析デザイン。
  - 識別戦略 (identification strategy) と呼ぶことも。
- 直観的な言い方１：コントロール・トリートメントグループを**「上手」に比較**する方法。
- 直感的な言い方２：反実仮想を妥当な形で推定する方法。
- リサーチデザインは問い・背景・設定・データによって異なってくる。
  - 状況により、適用するツールに必要な仮定が満たされるか否かが異なる。

因果効果推定のための様々なアプローチ

- **ランダム化比較試験**：実験によって、処置を無作為に割り付ける。
- **回帰分析・マッチング**：観察される属性を用いて、セレクションを統制する。
- **パネルデータ、差の差分法**： 個々人の変化しない異質性を統制する。
- **回帰不連続デザイン**：処置が不連続に変化するところに着目する(例：地理的境界)
- **操作変数法**:：処置には影響するが、アウトカムには直接影響しない変数を活用する。
- **構造推定アプローチ**： 反実仮想をモデルから予測(生成)する。

自然実験(Natural Experiment)

- 自然実験：トリートメントの有無があたかも「実験」されているかのようにランダムに決まる。
  - 政策変更、「たまたま」のイベント、政策適用の閾値、などなど。
- 自然実験があるとRCTの状況に近づき、分析がしやすくなる。
- 近年のミクロ実証分析では**自然実験 (Natural Experiment)**を活用することが重視されている。
  - 常にあるとは限らないが、思った以上に多い。
  - そして、どの程度「ランダムか」もケース・バイ・ケース。

具体的な手法・アプローチ

手法１：ランダム化比較試験 (RCT)

- 実験によって、興味ある変数の割当を無作為（ランダム）に割り当てる。
- 例：電力消費実験 (Ito et al 2018)

$$\log x_{it} = \beta M_{it} + \gamma E_{it} + \theta_i + \lambda_t + \eta_{it}$$

- 道徳的説得 M_{it} と経済インセンティブ E_{it} はランダムに決まっている。
- すなわち、誤差項 η_{it} と無相関 -> OLS でOK!!
- 問題点：RCTは常にできるとは限らない。

手法２：回帰分析・マッチング

- Selection on observable という仮定に基づく：処置の割当は観測される変数によってもたらされる。
- アイデア：同じ属性 xを持っている、「処置を受けた人」と「受けていない人」を比較すればOK
- 線形回帰分析においては、処置の割当に影響する変数 x_i をコントロールする。

$$y_i = \alpha_0 + \alpha D_i + \beta x_i + \epsilon_i$$

- 適切にコントロールすれば、欠落変数バイアスを回避できる。-> α を推定可能。

多くの変数をコントロールすれば良いのか？

- 多くの変数を入れると、興味ある変数との(不完全な)多重共線性の問題が生じる
  - 興味ある変数の係数推定値の標準誤差が大きくなる。
- 欠落変数バイアスを避けるために必要な変数が観察されないかもしれない。
  - 例：個々人の能力

因果効果推定の観点からみた線形回帰

- 以下の回帰式を用いて、 D の Y に対する因果効果を推定しよう。

$$y_i = \alpha_0 + \alpha_1 D_i + \beta' x_i + \epsilon_i$$

- **興味ある変数 D の変動(variation)** が以下の２点の意味で重要となる。

- 1: x_i を コントロールした後に、 **D が外生的な変動** を持つこと。
  - 仮定３（条件付き期待値ゼロ）と関連する。-> バイアスなし。

- 2: x_i を コントロールした後に、 **D が十分な変動**を持つこと。
  - 精緻に推定する上で重要 (小さい標準誤差)
  - 仮定４（多重共線性）と関連する。

(補足) 各種マッチング推定手法について

- 説明変数をコントロールした線形回帰モデルはマッチングアプローチの一種。
- 線形回帰モデルを用いないマッチングの方法も多数ある。
  - 傾向スコアマッチング
  - Nearest neighborhood matching
- 詳しくは各種教科書を参照。
- 実践上は線形回帰モデルが持ちいられることが非常に多い。

手法３：パネルデータ

- パネルデータがあれば(1)個人の異質性(個人固定効果)そして(2)時間固定効果をコントロール可能

$$y_{it} = \alpha_0 + \alpha D_{it} + \beta x_{it} + \epsilon_i + \epsilon_t + \epsilon_{it}$$

- データとして見えていないもの（例：能力）も個人固定効果として捉えられる。
- 仮定：個人・時間固定効果をコントロールした上で、説明変数 D_{it} が 誤差項 ϵ_{it} と無相関
- 注意点：時間を通じて変化しない変数の係数は推定できない。(多重共線性)

パネルデータの応用：差の差分法(DID)

- 回帰式：

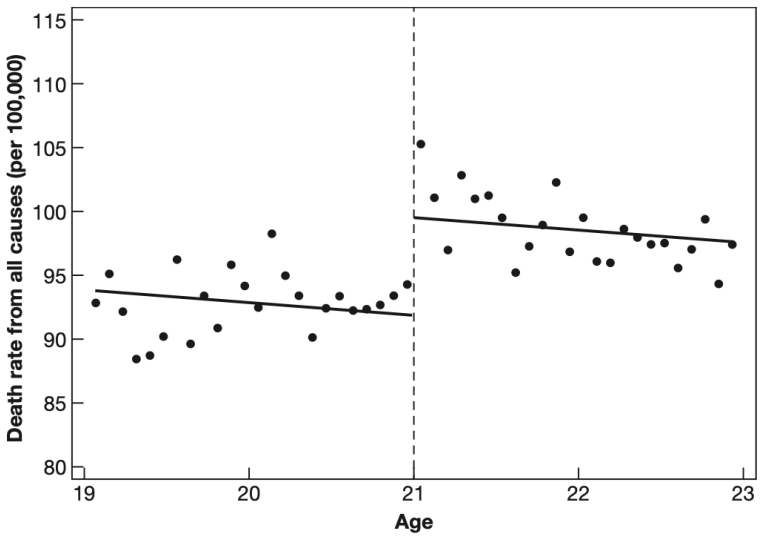
$$y_{it} = \alpha_0 + \alpha \cdot treat_i * after_t + \beta x_{it} + \epsilon_i + \epsilon_t + \epsilon_{it}$$

- DIDの並行トレンドの仮定は、(実質的に) $treat_i * after_t$ と ϵ_{it} が無相関であることと同一
- たくさんの例！
- 大阪大学国際公共政策大学院のファカルティによるDIDマニュアル

手法４：回帰不連続デザイン

- 処置がギリギリで変化する閾値の周りに着目する。
- 例１：UBERのsurge pricing
- 例２：法定飲酒可能年齢と死亡率

FIGURE 4.2
A sharp RD estimate of MLDA mortality effects



Notes: This figure plots death rates from all causes against age in months. The lines in the figure show fitted values from a regression of death rates on an over-21 dummy and age in months (the vertical dashed line indicates the minimum legal drinking age (MLDA) cutoff).

手法５：操作変数法

- モデル

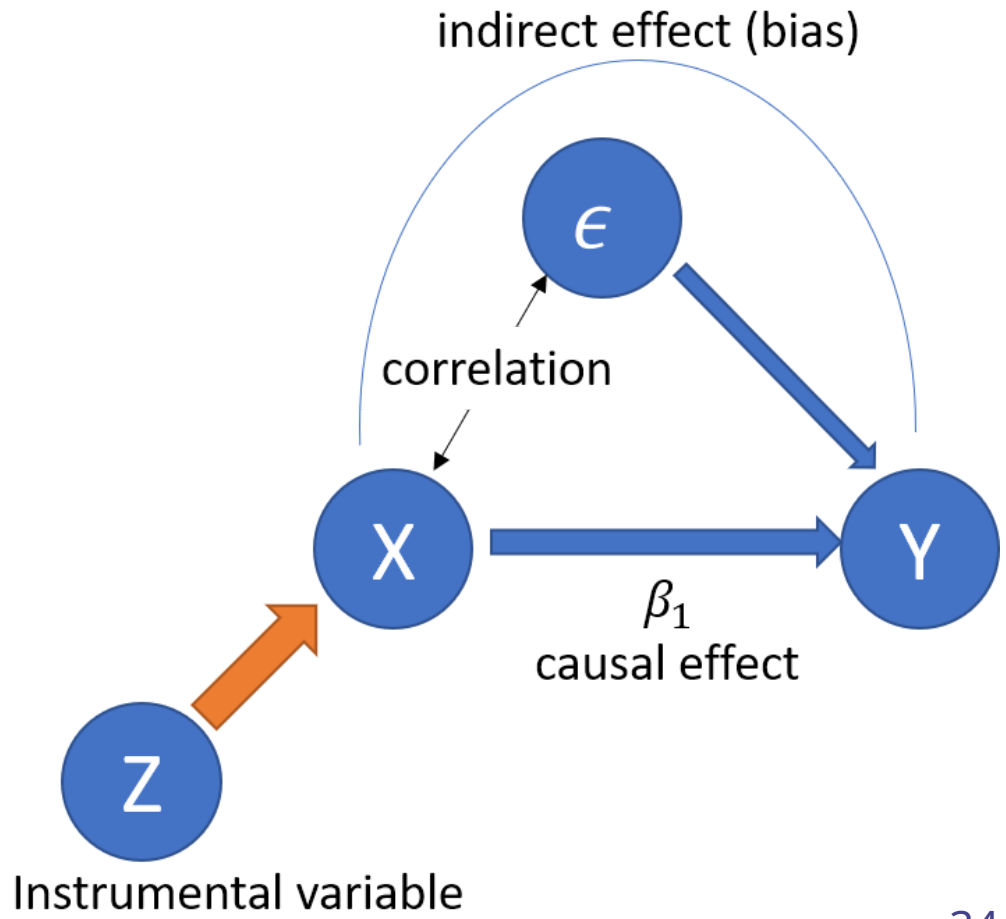
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ Cov(x_i, \epsilon_i) \neq 0$$

- 変数 z_i は **操作変数（instrumental variable; IV）** とよばれ、次の条件を満たす。
  1. **独立性（independence）**：$Cov(z_i, \epsilon_i) = 0$（操作変数と誤差に相関がない）
  2. **関連性（relevance）**：$Cov(z_i, x_i) \neq 0$（操作変数と内生変数に相関がある）

- アイデア：**操作変数 z_i に引き起こされる**内生変数 x_i の変動を用いて、 x_i の y_i に対する直接的な因果効果（ β_1 ）を推定する！

アイデア：操作変数があると・・・

- 操作変数 z を考える。
- 今、操作変数 z が動くと、 x が動くとしよう。(**関連性**の仮定)
- 一方、操作変数 z が動いても、誤差項 ϵ は変化しない (**独立性**の仮定)
- 操作変数 z によって引き起こされた x の変化が y とどのように相関するかを見ることによって、因果効果 β_1 を取り出すことができる！



手法６：構造推定

- 経済モデルの推定＆シミュレーション分析に基づく、因果効果の推定方法。
- 例：需要曲線と供給曲線
  - 需要曲線と供給曲線をデータから推定する。
  - 税率が変化したらどうなるか？
  - 企業同士が合併したらどうなるか？（合併シミュレーション)
- (ちなみに私の研究の８－９割は構造推定アプローチ)

実践上のアドバイス

効果検証型研究を行う上でのコツ

- 大きく分けて 2 つのアプローチ。
- リサーチデザイン・ベース：最初から「自然実験」を探す。
  - アカデミックリサーチではよくある。一方で若干の本末転倒感？
- クエスチョン・ベース
  - 分析したいアウトカム変数 Y と処置変数 X を決める。
  - Y と X の決定要因について考える。
  - 共通の要因がデータとして観察されるならば追加変数としてコントロールしよう。
  - 自然実験的なものがありそうならば、それを用いよう。

クエスチョンベースについてもう少し詳しく

- 綺麗なリサーチデザインや自然実験がない場合は、**「変数をコントロールした回帰分析からどこまで言えるか？」**を丁寧に検討・議論していくことになる。

- 例えば：
  - どのような変数が欠落しているか？
  - 欠落変数バイアスの方向性はどちらか？
  - コントロールする変数によって、推定値はどの程度変わるか？（頑健性の議論）

最後に: 責任感を持った分析が大事！

- 計量分析によって、何を見たい（推定したい）のか？
- その見たいものは、誰にとって意味がある・重要なものか？（分析のモチベーション)
- 現状の推定方法・識別戦略によって、それはどの程度うまく捉えられていそうか？
- 識別戦略のみならず、分析に用いる変数自体も重要。(適切な計測なのか？)