統計的機械学習　（応用計量分析２）第9回

未観測交絡因子への対処法（参考pdf 12章）

言語モデルの知識を用いる手法

データだけから因果構造全体を復元するのは難しい問題
→ 知識を言語モデルから抽出する

- データのみから因果構造全体を復元することはかなり難しく、実用的には人間がある程度の構造を事前知識として与えることが多い
- 「（変数名A）を変更すると（変数名B）に変化が生じますか？」と大規模言語モデルに質問
  - 変数名が存在すること、その因果関係が一般的知識として知られていることが前提
    - 大規模言語モデルにその一般的知識が学習され埋め込まれている前提
  - 変数ペアの因果方向判定タスクで既存のデータに基づく手法を凌駕
- ただし新たな知識構築（科学の発展）にはデータも必要

Model	Acc.	Wt. Acc.
Slope (Marx & Vreeken, 2017)	0.75	0.83
bQCD (Tagasovska et al., 2020)	0.68	0.75
PNL-MLP (Zhang & Hyvarinen, 2012)	0.75	0.73
Mosaic (Wu & Fukumizu, 2020)	0.83	0.82
ada	0.50	0.50
text-ada-001	0.49	0.50
babbage	0.51	0.50
text-babbage-001	0.50	0.50
curie	0.51	0.52
text-curie-001	0.50	0.50
davinci	0.48	0.47
text-davinci-001	0.50	0.50
text-davinci-002	0.79	0.79
text-davinci-003	0.82	0.83
LMPrior (Choi et al., 2022)	0.83	-
gpt-3.5-turbo	0.81	0.83
gpt-3.5-turbo (causal agent)	0.86	0.87
gpt-3.5-turbo (single prompt)	0.89	0.92
gpt-4 (single prompt)	**0.96**	**0.97**

Kiciman, Emre, et al. "Causal reasoning and large language models: Opening a new frontier for causality." Transactions on Machine Learning Research (2023).

振り返り

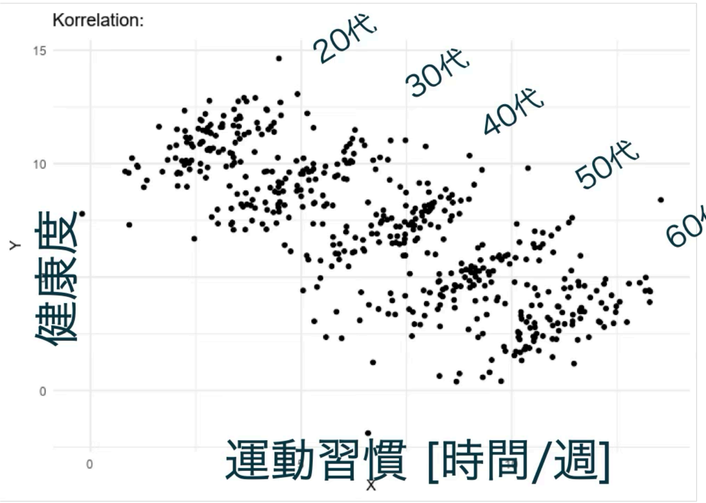
因果構造を推定する問題＝因果探索
仮定の置き方と得られる結論の強さにより幾つかのアプローチ

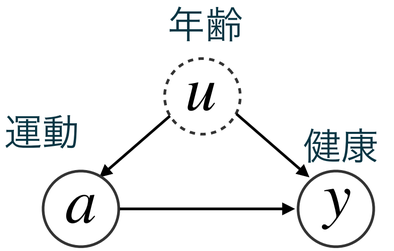
- 観察データのみから因果構造を完全に特定することは不可能
  - 一部の辺の向きの識別は諦めるか、追加でモデル等に仮定を置くことで識別する
- IC/PC法
  - （条件付き）独立性のみに基づく辺の枝刈りとV構造を用いた向き付け
  - 仮定が少ないが、向きが判定できない辺が残る可能性がある（CPDAGまで識別）
- LiNGAM
  - 線形性を仮定、ノイズ分布はガウス分布でないことを仮定
- NOTEARS
  - 構造方程式 f のモデルをノイズ含めて仮定
  - トレース指数関数により非巡回制約を連続空間で表現
- 言語モデルからの知識抽出
  - 変数ペアの因果関係を問い合わせることで知識を抽出

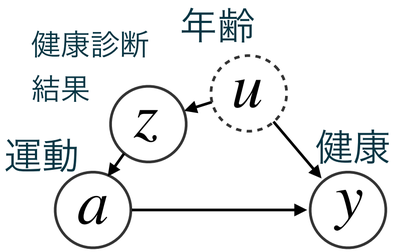
出欠コード
XXXX

未観測交絡因子の問題は重大

- 交絡因子（u）が観測できていなかったら
  一般に調整できない
  - 無視可能性を満たすには交絡因子の観測が必要
  - $(y_{a'})_{a' \in \mathcal{A}} \perp\!\!\!\perp a \mid u$
- ただし、バックドア基準を満たせる場合もある
  - 右下図の健康診断結果 z を調整することでも代替可
- 現実の問題では交絡因子の候補も不明なことも
  - データだけで未観測交絡因子の不存在を確認する術はない
    - ※ LiNGAMを仮定すると未観測交絡因子の”検出”は可能な場合も
- そのような場合にも識別可能な方法は？



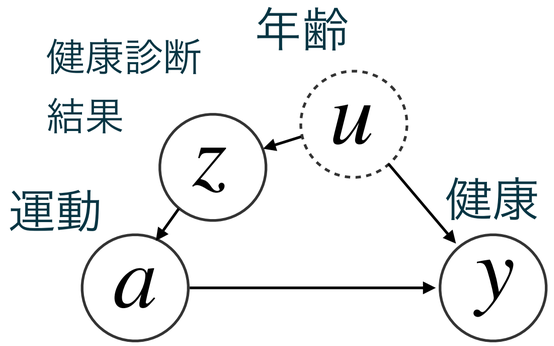




バックドア基準を用いて確認

因果構造の情報を用いて調整する変数集合を選ぶ基準

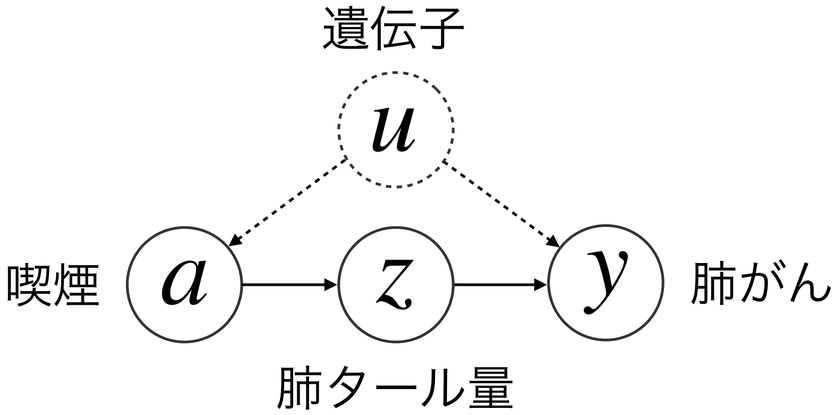
- 健康診断結果 z はバックドア基準を満たす
  - バックドアパスの鎖上の変数を用いてブロック
- バックドア基準（再掲）
  - 因果ダイアグラムGにおいて、aからyへの有向パスがあるとする。次の２条件を満たすとき、**変数集合 Z は順序対 (a, y) についてバックドア基準を満たす**という
    - (B1) aからZへの有向パスがない（行動より下流の変数を含まない）
    - (B2) aに入るパスを含む、aとy を結ぶパス（**バックドアパス**）において、ZがaとYを**有向分離** （ブロック）する
  - ただし、a-y間の全てのパス p に対してZが以下の条件のいずれかを満たすとき、Zはaとyを**有向分離する**という
    - **鎖 i→m→j** またはフォーク i←m→j を含み、m は Z に含まれる
    - 合流点 i→m←j を含み、m及びその子孫は Z に含まれない



フロントドア調整とは

フロントドア調整は因果パスを組み合わせるバックドア基準の応用

- 例：喫煙→肺がんの因果機序の論争（1950—‘60年代）
  - 「喫煙を欲させる遺伝子uがある」主張
    - ※実際には、リスク増加率が大きく遺伝子だけでは説明つかないことなどで結論
- 因果機序を分解する
  - 喫煙 a →肺のタール沈着量 z →肺がん y
  - 遺伝子→肺のタール沈着量とは考えづらい
- $a \to z$ と $z \to y$ はそれぞれ識別可能
  - $a \to z$ はバックドアパス無し
  - $z \to y$ はバックドアパスがあるが、aが有向分離する
- **媒介変数** z を介した因果機序を推定して組み合わせれば
  $a \to y$ の因果効果ATEは識別可能



フロントドア調整の定理

フロントドア基準が満たされればフロントドア調整によってATEが識別される

- フロントドア基準
  - 変数の集合zが順序対(a, y)に関して**フロントドア基準**を満たすとは、以下の3つの条件を満たすこと
    1. zはaからyへのすべての有向パスを遮断する
    2. aからzへのバックドアパスが存在しない
    3. zからyへのすべてのバックドアパスがaによってブロックされる
- zがフロントドア基準を満たすとき、$a \to z$ と $z \to y$ は以下

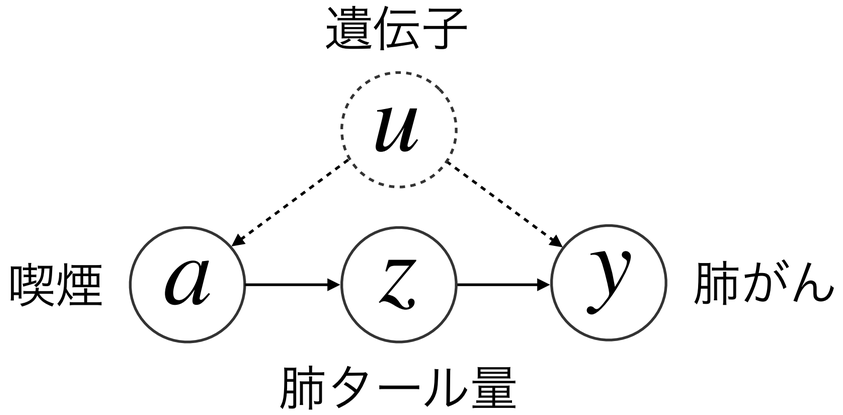
$$p(z \mid \mathrm{do}(a)) = p(z \mid a)$$

  - $$p(y \mid \mathrm{do}(z)) = \sum_{a} p(y \mid a, z) p(a)$$

- これらを組み合わせると以下が成立

$$p(y \mid \mathrm{do}(a)) = \sum_{z} p(z \mid \mathrm{do}(a)) p(y \mid \mathrm{do}(z))$$

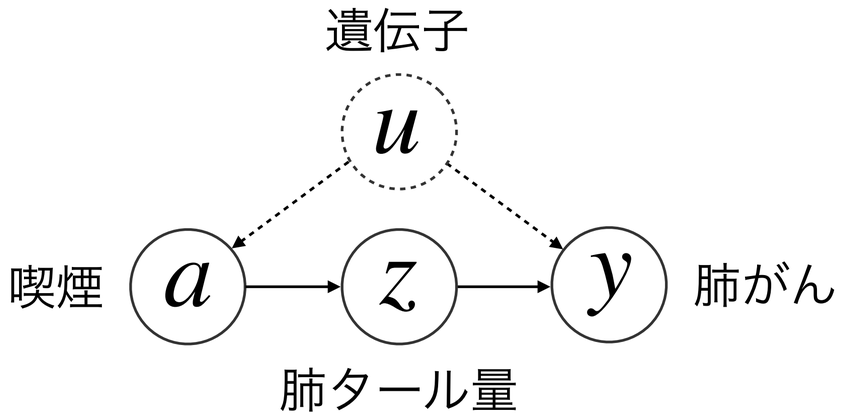
  - $$= \sum_{z} p(z \mid a) \sum_{a'} p(y \mid a', z) p(a')$$



フロントドア調整の具体的方法（線形モデル）

$a \to z$と$z \to y$をそれぞれ推定して総合する

- 線形モデルを仮定
  - $\mathbb{E}[y \mid a, u] = \beta'_{y0} + \beta_{ay} a + \beta_{uy} u$ —（★）
    - $\beta_{uy} u$の項が邪魔
- 分解する
  - $\mathbb{E}[z \mid a] = \beta_{z0} + \beta_{az} a$
  - $\mathbb{E}[y \mid z, a] = \beta_{y0} + \beta_{zy} z + \beta'_{ay} a$ ・・・推定可能
- 式展開
  - $$\begin{aligned}\mathbb{E}[y \mid a, u] &= \mathbb{E}_z[\underbrace{\mathbb{E}[y \mid z, a, u]}_{=\mathbb{E}[y \mid z, u]} \mid a, u] \\ &= \mathbb{E}_z[\beta''_{y0} + \beta_{zy} z + \beta_{uy} u \mid a, u] \\ &= \beta''_{y0} + \beta_{zy} \underbrace{\mathbb{E}[z \mid a, u]}_{=\mathbb{E}[z \mid a]} + \beta_{uy} u \\ &= \beta''_{y0} + \beta_{zy}(\beta_{z0} + \beta_{az} a) + \beta_{uy} u\end{aligned}$$
  - （★）と比較して、$\beta_{ay} = \beta_{zy}\beta_{az}$ → 係数をそれぞれ推定して積をとればよい

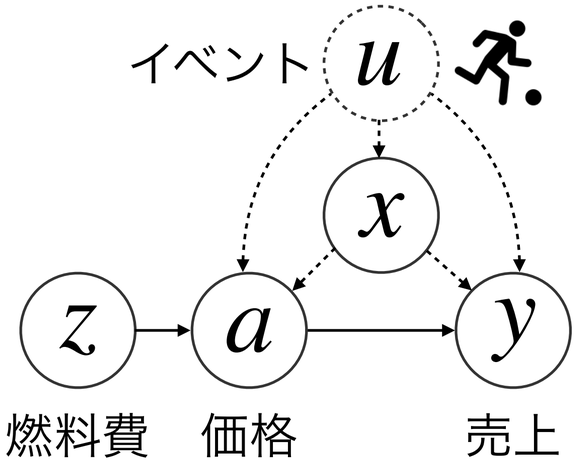


操作変数（Instrumental Variable；IV）法とは

出欠コード
XXXX

独立した変数（外生変数）を弱いクジとしてその影響を分析

- 例：航空機チケット価格弾力性
  - 近隣でイベントuがあると売上が上昇
    - イベント情報は得られないとする
  - 早期から売れ行きが良いのでアルゴリズムにより価格設定が上昇
- 交絡因子と独立した行動決定要因zを仮定
  - 燃料費（原油価格）はイベントと独立 $z \perp\!\!\!\perp u$ かつ
    チケット価格を通してしか売上に影響しない（**除外制約**）
  - 燃料費z→売上yの因果効果は価格a→売上yの因果を通ったもの
  - 燃料費z→価格aの因果効果と併せて考えれば価格a→売上yがある程度わかる
- 2021年ノーベル経済学賞はIV法関連
  - IV法を用いて教育効果を分析したカード
  - IV法の理論的基盤を整備したアングリスト・インベンス



Ill. Niklas Elmehed © Nobel Prize Outreach.

デイビッド・カード教授

Ill. Niklas Elmehed © Nobel Prize Outreach.

ヨシュア・アングリスト教授

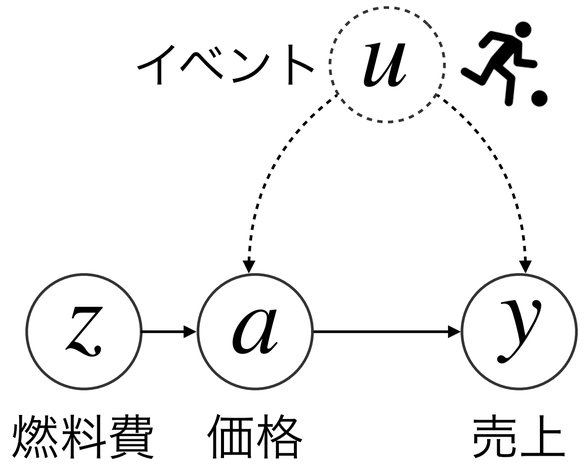
Ill. Niklas Elmehed © Nobel Prize Outreach.

グイド・インベンス教授

操作変数法（線形モデル）

$z \to a$と$z \to y$の分解を用いて式展開

- 線形モデルを仮定
  - $\mathbb{E}[y \mid a, u] = \beta_{y0} + \beta_{ay} a + \beta_{uy} u$
    - uを条件に入れればバックドア基準を満たし
      β_{ay}が求めたい因果効果。ただし$\beta_{uy} u$の項が邪魔
  - 簡単のためxは無いとする
- 操作変数zからa, yへの因果効果
  - $\mathbb{E}[a \mid z] = \beta_{a0} + \beta_{za} z$
  - $\mathbb{E}[y \mid z] = \beta'_{y0} + \beta_{zy} z$
- ここで、$z \to y$を$z \to a$と$a \to y$に分解して表式
  - $\mathbb{E}[y \mid z] = \mathbb{E}_{a,u}[\mathbb{E}[y \mid a, z, u] \mid z] = \mathbb{E}_{a,u}[\mathbb{E}[y \mid a, u] \mid z]$
    - 2つ目の等式はzとuの独立性から従う



係数の比（変数数が同じ）か２段階最小二乗法

- 式展開

$$
\begin{aligned}
\mathbb{E}[y\,|\,z] &= \mathbb{E}_{a,u}[\mathbb{E}[y\,|\,a,u]\,|\,z] \\
&= \beta_{y0} + \beta_{ay}\mathbb{E}[a\,|\,z] + \beta_{uy}\underbrace{\mathbb{E}[u\,|\,z]}_{=\mathbb{E}[u]} && \because\ \mathbb{E}[y\,|\,a,u] = \beta_{y0} + \beta_{ay}a + \beta_{uy}u \\
&= \beta''_{y0} + \beta_{ay}\mathbb{E}[a\,|\,z] && \leftarrow\ (\bigstar)\ \ \beta''_{y0} := \beta_{y0} + \beta_{uy}\mathbb{E}[u] \\
&= \beta''_{y0} + \beta_{ay}\left(\beta_{a0} + \beta_{za}z\right) && \because\ \mathbb{E}[a\,|\,z] = \beta_{a0} + \beta_{za}z \\
&= \beta'''_{y0} + \beta_{ay}\beta_{za}z
\end{aligned}
$$

- モデル $\mathbb{E}[y\,|\,z] = \beta'_{y0} + \beta_{zy}z$ と見比べると
  - $\beta_{zy} = \beta_{ay}\beta_{za} \Rightarrow \beta_{ay} = \beta_{zy}/\beta_{za}$
  - a, z共に1変数の場合は推定した係数の比を取るだけで推定可
- より一般には、$\mathbb{E}[a\,|\,z]$を回帰で予測した上で(★)に代入する**二段階最小二乗法**

操作変数法の推定対象

操作変数法はATEを識別しない
単調性を追加で仮定すれば局所因果効果LATEを識別する

- 線形モデルが正しければβ_{ay}はATEを識別する
  - $\mathbb{E}[y \mid a, u] = \beta_{y0} + \beta_{ay}a + \beta_{uy}u$
- しかしモデルが不正確な場合はそうはいえない
- 追加で**単調性**を仮定する
  - 各ユニットiにおいて、z^iの増加に対してa^iは非減少
    - 行動をzによって「促された」ら、（反応しないことはあっても）
      「逆の行動をとる」あまのじゃく（非遵守者；Defiers）は居ない
- このとき、IV法は遵守者（Compliers）に限定した平均因果効果を識別
  - $c^i = 1 \Leftrightarrow a^i_{z=1} > a^i_{z=0}$ と定義して
  - LATE $= \mathbb{E}[y_{a=1} - y_{a=0} \mid c^i = 1]$
- 行政施策等（z）を打った際に、それに反応した人だけに限定した効果と解釈できる

まとめ

未観測交絡因子によりバックドア基準を満たせない場合の手法を2つ紹介
フロントドア調整は媒介変数で分解して総合、操作変数法は外生変数をランダマイザとみなす

- 未観測交絡因子があってもバックドア基準を満たせる場合もある
- どうしても満たせない場合、または未観測交絡の疑いがいくらでも出てくる場合は別のアプローチも考えられる
- 2つのアプローチを紹介
  - フロントドア調整は未観測交絡因子から影響を受けていない、かつ$a \to y$の因果効果を完全に媒介する変数を用いて分解して総合
  - 操作変数法は未観測交絡因子と独立かつaを通してしかyに影響しない（かつaには影響する）変数をクジ（ランダマイザ）として、その影響によるaの変動のyへの伝播を分析