

2022年3月26日
第24回春の合宿セミナー（日本行動計量学会）
（統計的因果推論入門）

講義1

統計的因果推論の基礎の基礎

長崎大学 情報データ科学部 准教授

高橋 将宜

博士（理工学）

m-takahashi@nagasaki-u.ac.jp

概要

- 自己紹介
- シラバスの確認
- 白髪の人には禿げない？
- ごま油と長寿の関係
- 新型コロナとワクチンの関係
- チョコレートの消費量とノーベル賞受賞者数との関係
- 統計的因果推論の入り口



自己紹介

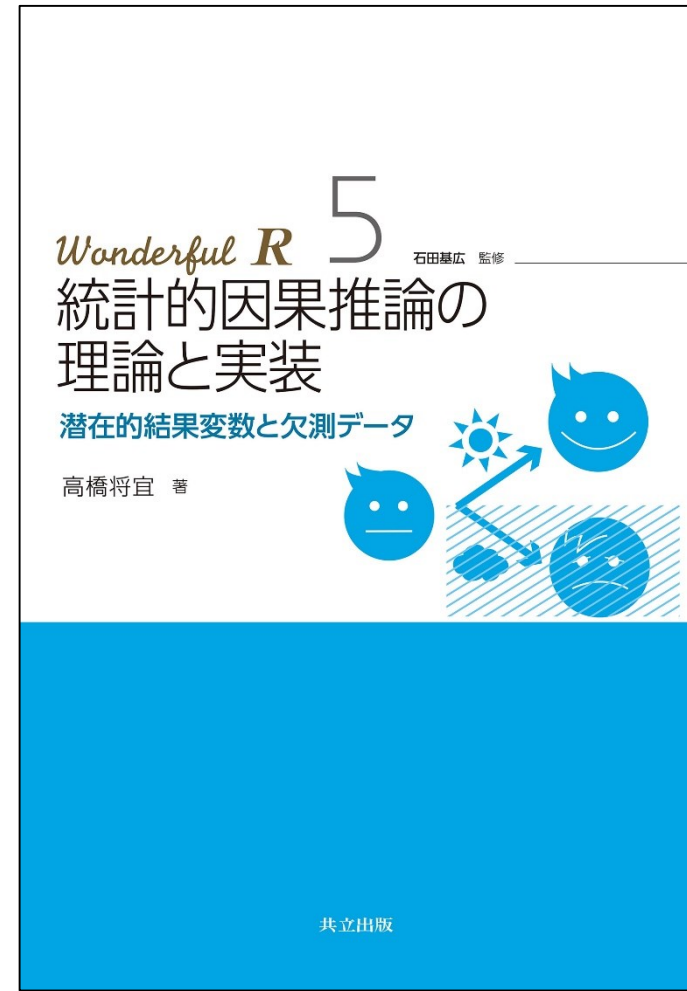
職位・専門・学位など

- 名前：高橋 将宜（たかはし まさよし）
- 現職：長崎大学 情報データ科学部 准教授
- 専門
 - 統計科学，計量政治学
 - 欠測データ解析，統計的因果推論
- 学位
 - 博士（理工学）成蹊大学理工学研究科情報科学コース
 - 修士（政治学）ミシガン州立大学
- 資格
 - 専門統計調査士，英検1級，TOEIC 990点（満点）

著書



『欠測データ処理：Rによる単一代入法と多重代入法』
(2017年，共立出版)



『統計的因果推論の理論と実装』
(2022年，共立出版)

最新の研究： MIRDDという新たな統計的因果推論の手法を提案



国立大学法人
長崎大学
 NAGASAKI UNIVERSITY

> サイトマップ > お問い合わせ > 取材・撮影申込 > アクセ






受験生
 在学生
 卒業生
 保護者
 地域

大学案内
 学部・大学院等
 教育・学生生活
 研究・産学官連携
 留学・国際

HOME > Research > 詳細



Research

2021年08月23日

新たな統計的因果推論手法の提案（統計学の新知見）

たとえば、薬の効果を知りたいとき、「薬を飲んだ場合の結果」と「薬を飲まなかった場合の結果」は同時に調べることができません。あらゆる科学において、因果推論は研究の重要な目標ですが、このような潜在的な結果の差から考える必要があるにも関わらず、潜在的結果のうちの片方しか観測されないことから、統計的因果推論は「欠測データ（データの一部が観測されない）問題」といわれています。

欠測データの対処法として多重代入法が知られていますが、閾（しきい）値における局所的な平均因果効果を推定する手法として、多重代入法を活用することは、これまで議論されて来ませんでした。

そこで、情報データ科学部 の高橋 将宜准教授は、多重代入法によって閾値における局所的な平均因果効果を適切に推定できることを示しました。また、フリーソフトRにより分析ツールを開発して、簡便に利用できる環境も整えました。

本研究で得られた成果は、インパクトファクター付の統計学専門誌「Communications in Statistics - Simulation and Computation(Taylor & Francis)」に掲載されました。応用例として、「選挙における既存勢力の優位性に関する分析」、「教育と貧困の関係に関する分析」、「新型コロナワクチン接種に関する分析」など、今後、因果推論を目的とするさまざまな研究に寄与することが期待されます。

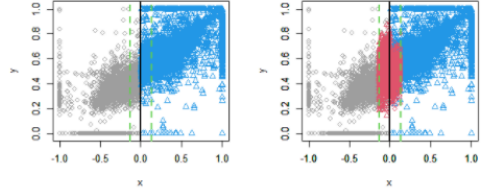


図1：観測データ（左図）と潜在的結果のシミュレーション（右図）

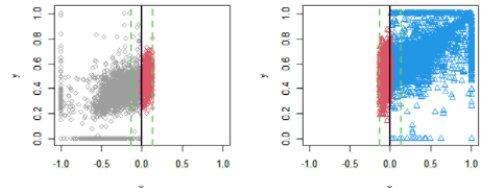


図2：非処置群の観測データとシミュレーション（左図）と処置群の観測データとシミュレーション（右図）

凡例：

- 黒縦線：閾値
- 緑縦破線：局所範囲
- 灰色の○：非処置群の観測値
- 青色の△：処置群の観測値
- 赤色の○：非処置群のシミュレーション値
- 赤色の△：処置群のシミュレーション値

■論文タイトル

Multiple imputation regression discontinuity designs: Alternative to regression discontinuity designs to estimate the local average treatment effect at the cutoff

（多重代入法回帰不連続デザイン：閾値における局所的な平均処置効果を推定するための回帰不連続デザインに代わる手法）

■掲載誌

Communications in Statistics - Simulation and Computation(Taylor & Francis)

<https://doi.org/10.1080/03610918.2021.1960374>

オンライン公開日：2021年8月18日（オープンアクセス）

■著者名

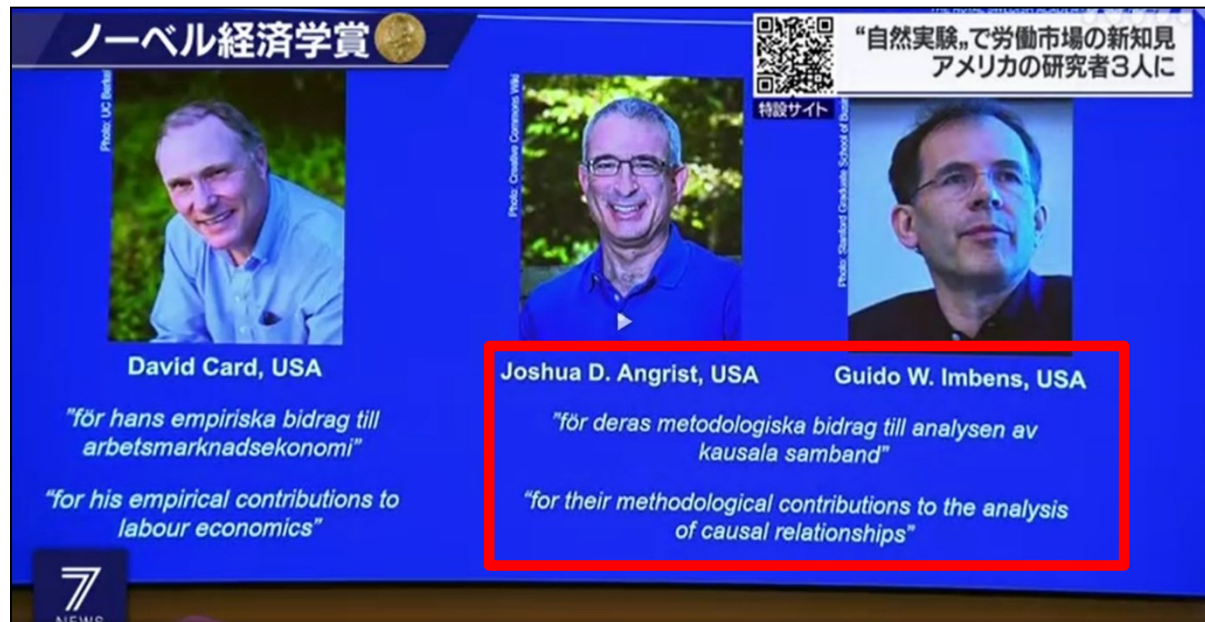
高橋 将宜：長崎大学情報データ科学部 准教授

https://www.idsci.nagasaki-u.ac.jp/research_ac/facultylist/staff19/

シラバスの確認

本コースの内容（1）

- 2021 年のノーベル経済学賞は、Joshua AngristとGuido Imbensという計量経済学者が受賞しました。
- その理由は、「因果関係の分析に対する方法論的な貢献に対して」でした。このように、統計的因果推論は、近年、世界的にさまざまな分野で注目されています。



本コースの内容（2）

□ 本コースでは

- 統計的因果推論の基礎的な考え方と応用的な技術について学びます.

□ 初日の講義

- 主に、統計的因果推論の考え方に焦点を当てています.
- ここで、**潜在的結果変数の枠組み**による因果推論の習慣を身に付けます.

本コースの内容（3）

□ 2日目の講義

- 主に、統計的因果推論の技術に焦点を当てています.
- **重回帰分析**で交絡因子を統制することの具体的な意味を理解した上で、その限界について認識するところから始めます.
- その後、**傾向スコアマッチング**、**操作変数法**、**回帰不連続デザイン**といった統計的因果推論でよく用いられる技術について、具体的に学びます.
- また、実際に統計環境Rで実行する方法についても学びます.

本コースの内容（4）

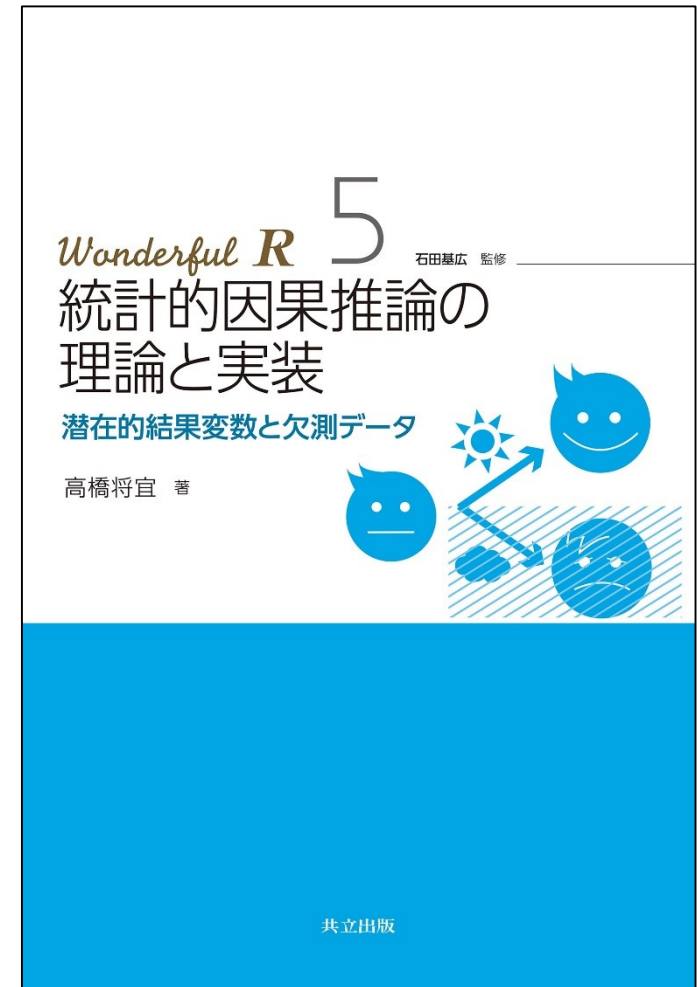
- 本コースでは、微分積分や線形代数を使わずに、数値計算とグラフによって解説をしますので、数学の難易度は比較的に低めに設定しています。

- 前提とする数学知識
 - 数I, 数A, 数II, 数B
 - ＊ただし、数Bのベクトルは使用しません。数IIの微分法と積分法は演算としては使用しませんが、極限や積分といった言葉は出てくるかもしれません。これらについては、もし使用する際には簡単な解説をする予定です。数IIIレベルは使いません。

本コースの内容（5）：教科書

- 講義内容は，高橋将宜
（2022）『統計的因果推論
の理論と実装』（共立出版）
に準拠しています．

- ISBN: 978-4-320-11245-2



初日

- 9:00～10:30 講義1（教科書Ch.1）
 - 統計的因果推論の基礎の基礎
- 10:40～12:10 講義2（教科書Ch.2）
 - 潜在的結果変数の枠組み
- 12:10～13:00 昼食休憩
- 13:00～14:30 講義3（教科書Ch.3）
 - 統計的因果推論における重要な仮定
- 14:50～16:20 講義4（教科書Ch.6）
 - 重回帰分析による交絡因子の統制の意味
- 16:30～17:00 質疑応答

2日目

- 9:00～10:30 講義5（教科書Chs.7-9）
 - 重回帰分析の限界と傾向スコアの導入
- 10:40～12:10 講義6（教科書Chs.10-11）
 - 傾向スコアの導入と傾向スコアマッチング
- 12:10～13:00 昼食休憩
- 13:00～14:30 講義7（教科書Chs.13-14）
 - 操作変数法の基礎
- 14:50～16:20 講義8（教科書Chs.15-18）
 - 回帰不連続デザインの基礎
- 16:30～17:00 質疑応答

統計的因果推論とは？

□ 因果関係とは？

- 要因Xを**操作する**とき， 要因Yが変化すること
- 操作なくして因果なし（no causation without manipulation）

□ 因果推論とは？

- 因果関係は目に見えるものではないため， 因果を推し測って考えるということ

□ 統計的因果推論とは？

- 因果推論をデータに基づいて行うもの

白髪の人には禿げない？

白髪の人と禿げの人

Aさん：白髪で禿げて
いない



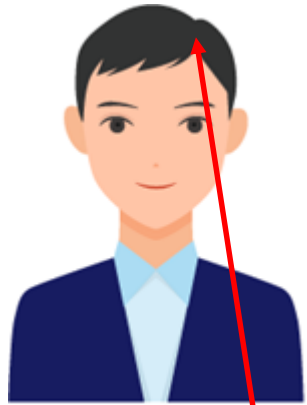
Bさん：禿げている



俗に「白髪の方は禿げない」という

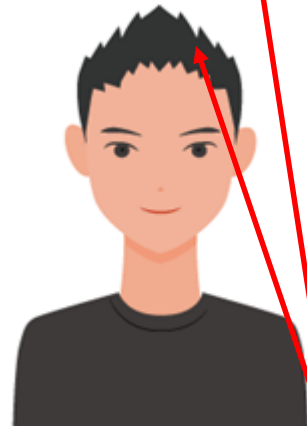
白髪の人には禿げない？

若い頃→老人



Aさん

加齢



Bさん

加齢



若いころには、誰しも髪があって黒い

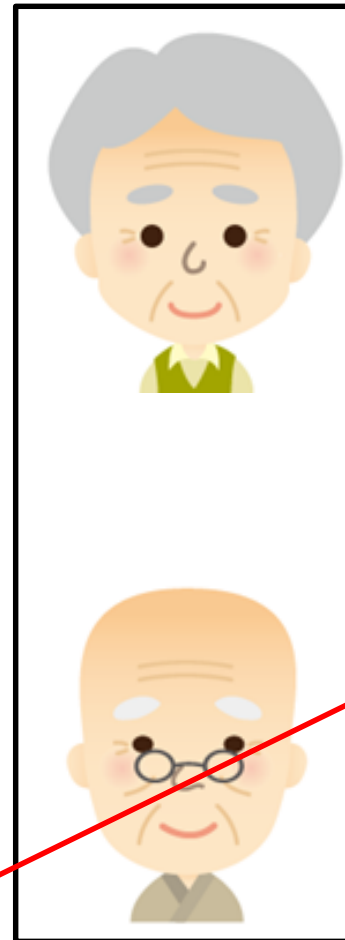
潜在的結果変数

白髪の人には禿げない？

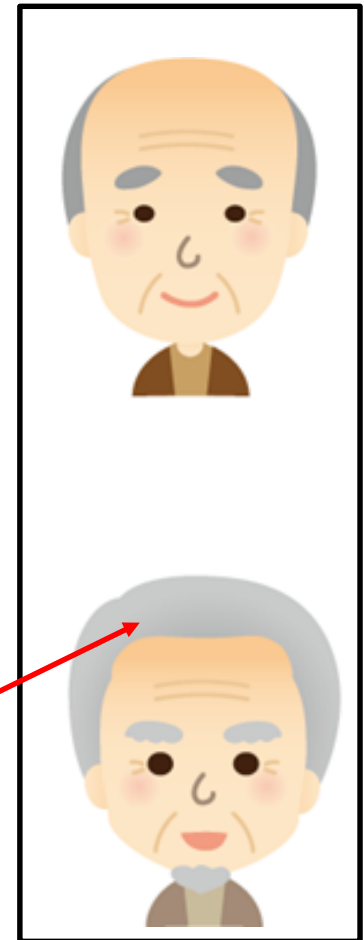
反実仮想：事実と反事実（仮定法）



Aさん



事実



反事実

黒か白か観測されない

潜在的結果変数による統計的因果推論

- 統計的因果推論を支える考え方は、観測されなかった「事実」について思いをはせる**反実仮想**です。
 - 反実仮想とは、英語で学ぶ**仮定法**のことです。



If I were a bird, I could fly.
(but I am not a bird, so I cannot fly).



ごま油と長寿の関係

テレビのバラエティ番組



原因？ → 結果？

- **ご長寿**は毎日の食事に**ごま油**を欠かさないという事例が紹介されているとしよう。
- **ごま油**を毎日欠かさないから、**ご長寿**なのだろうか？

条件付き確率の復習 (1)

Bを条件としたときのAの確率

$$Pr(A|B)$$

Aを条件としたときのBの確率

$$Pr(B|A)$$

$$Pr(A|B) \neq Pr(B|A)$$

一般的に、この2つの確率は一致しない

条件付き確率の復習 (2)

$$Pr(A|B) = \frac{Pr(B|A) \times Pr(A)}{Pr(B)}$$

■ $Pr(A) = Pr(B)$

■ $Pr(A|B) = Pr(B|A)$

本当に知りたい確率はどちら？

- ある人がご長寿であるとき, ごま油を摂取している確率

$$Pr(\text{ごま油を摂取} | \text{ご長寿})$$

- ある人がごま油を摂取していた場合に, ご長寿になれる確率

$$Pr(\text{ご長寿} | \text{ごま油を摂取})$$

本当に知りたい確率はこちらのはず

本来比較すべきこと：事実と反事実の比較

- ごま油を摂取していた場合に、長寿になる確率

$$Pr(\text{ご長寿} | \text{ごま油を摂取})$$

- ごま油を摂取していなかった場合に、長寿になる確率

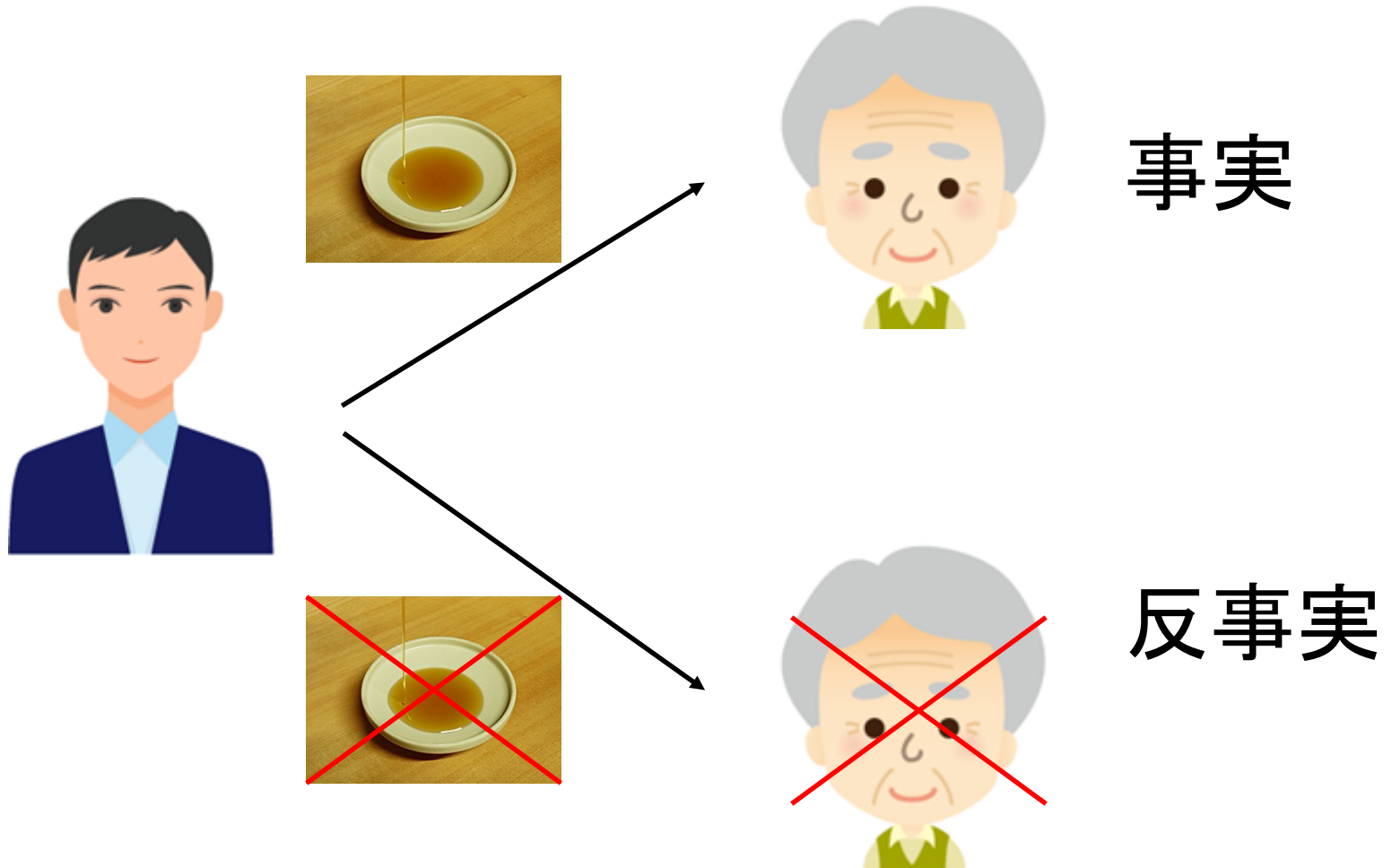
$$Pr(\text{ご長寿} | \text{ごま油を摂取しない})$$

$Pr(\text{ご長寿} | \text{ごま油を摂取}) > Pr(\text{ご長寿} | \text{ごま油を摂取しない})$ ならば、ごま油には寿命に対する因果効果があるかもしれない。

潜在的結果変数

ごま油と長寿の関係

〇〇していれば、△△でなかったのに（仮定法）



操作なくして因果なし

クロス集計表1

	長寿	短命
ごま油あり	81	?
ごま油なし	?	?

関連があるかどうか分からない

クロス集計表2

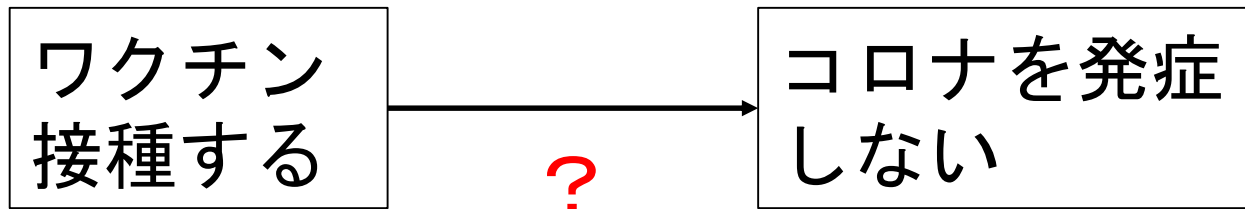
	長寿	短命
ごま油あり	81	19
ごま油なし	25	75

関連はありそう
因果があるかどうかは分からない

新型コロナとワクチンの関係

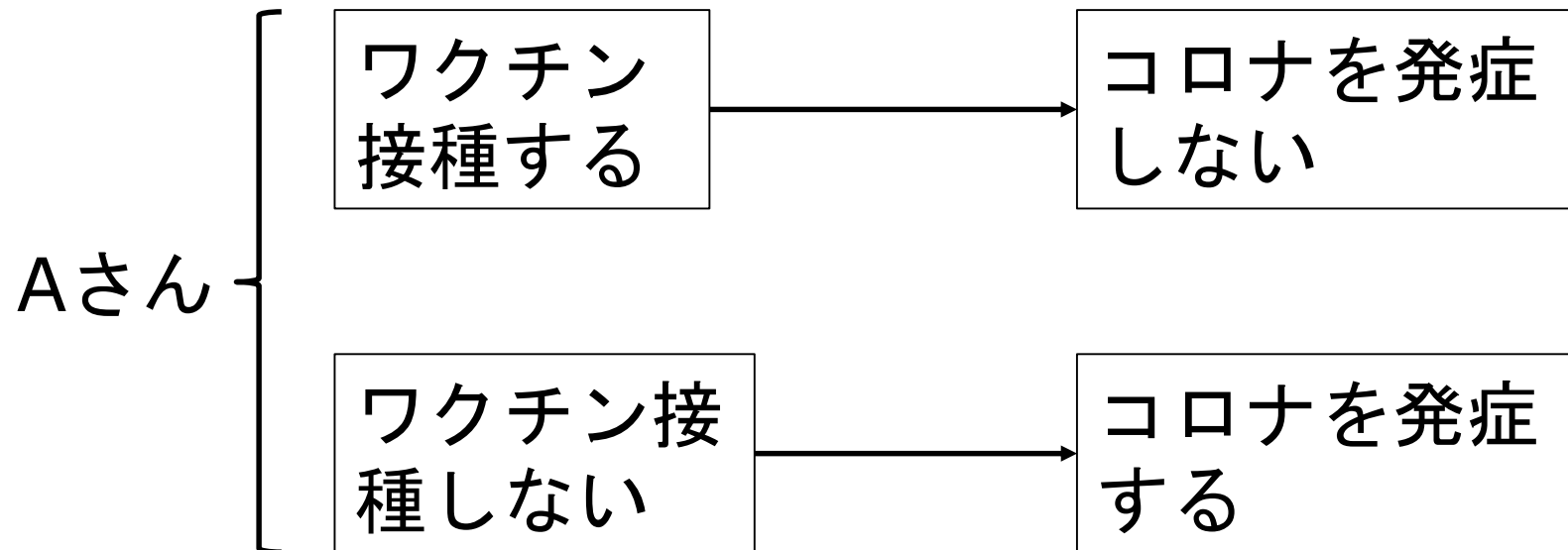
新型コロナとワクチン

- ある人がワクチンを接種したところ、新型コロナを発症しなかった。
- このワクチンには効果があると言えるか？



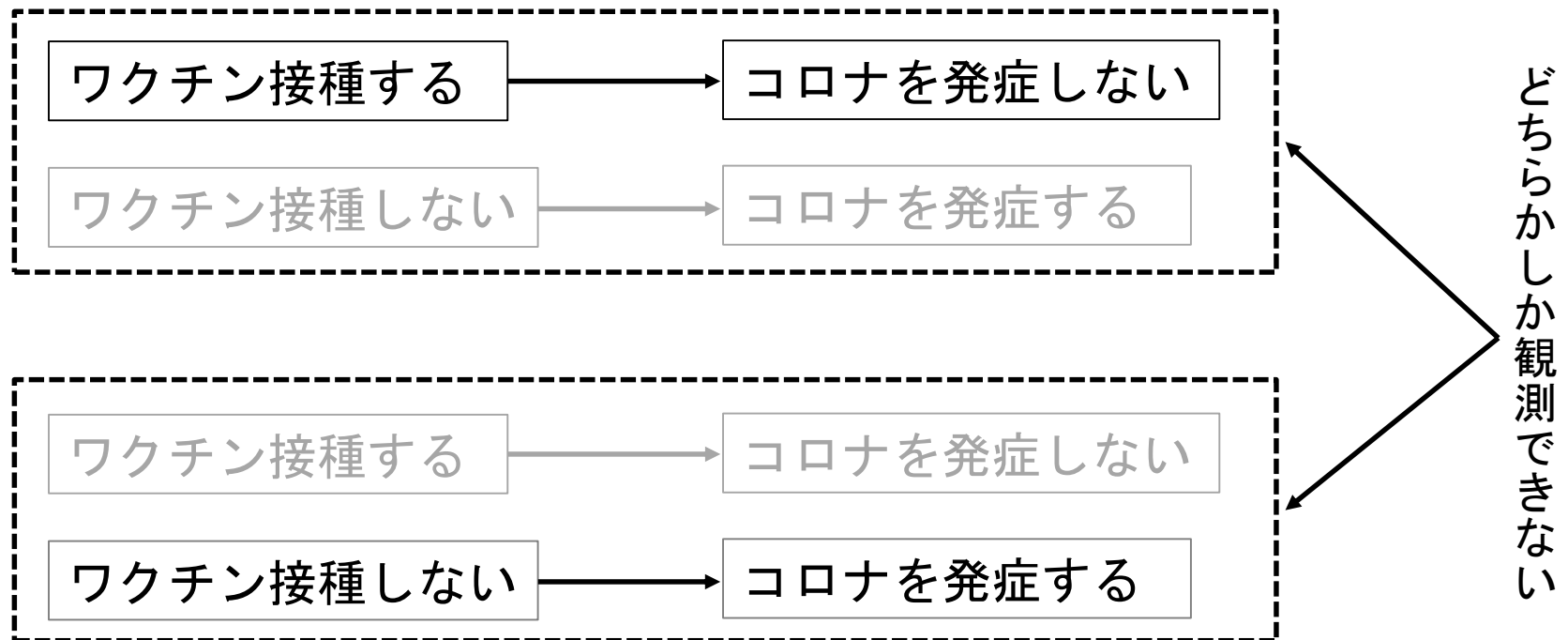
個体因果効果

- 同じ人について、ワクチンを接種した場合としなかった場合を比較して、新型コロナを発症するかどうかを比較する.
- このワクチンには効果があると言える.



因果推論の根本問題

- 同じ人について、ワクチンを接種した場合としなかった場合のどちらかしか観測されない。



個体因果効果は、定義はされても、観測も推定もできない。

因果推論は欠測データの問題

id	結果	処置	結果0	結果1
1	1	0	1	
2	1	0	1	
3	0	1		0
4	0	1		0
⋮	⋮	⋮	⋮	⋮
$n - 1$	1	1		1
n	0	0	0	



<https://www.kyoritsu-pub.co.jp/bookdetail/9784320112568>

無作為割付け：実験研究による平均因果効果の推定

被験者
36,523人

無作為に割り付け

処置群
18,198人
ワクチンを打つ

発症率0.04%

統制群
18,325人
ワクチンを打たない

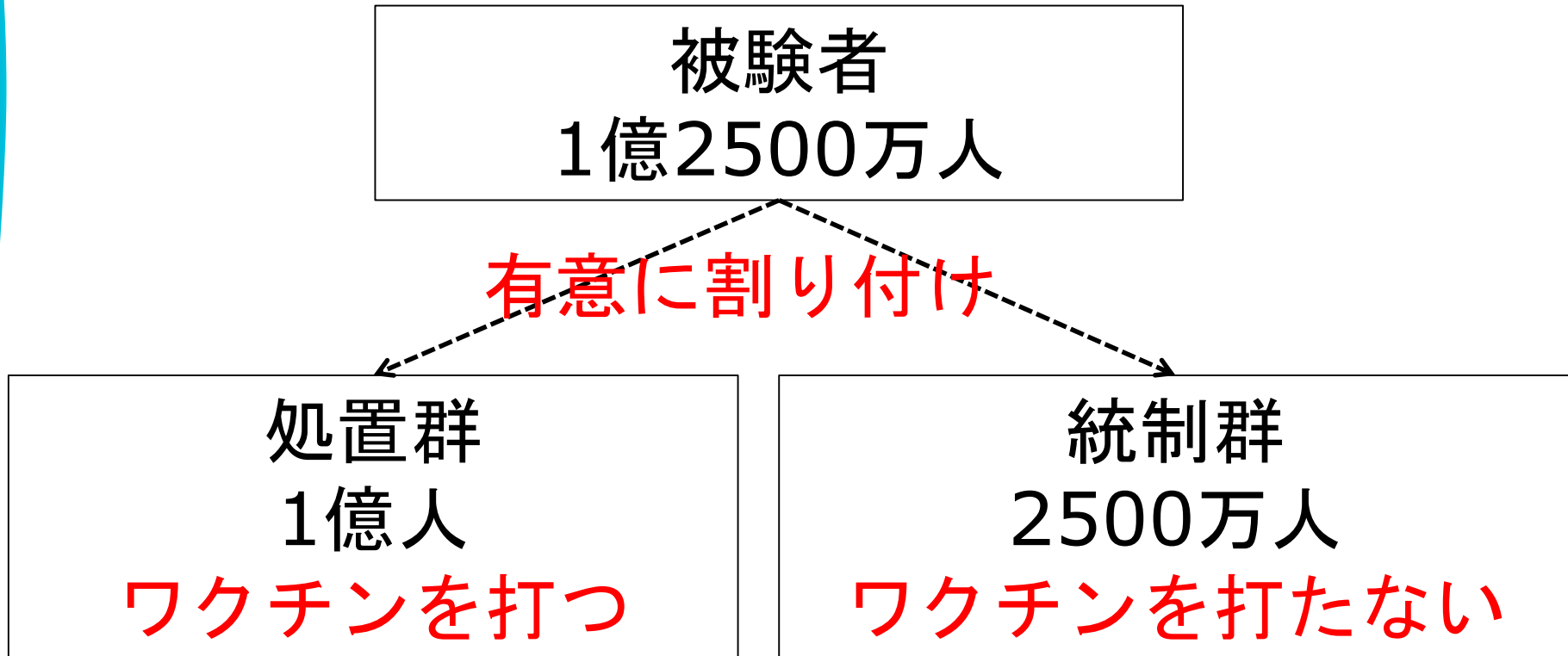
発症率0.88%

厚生労働省

ファイザー社の新型コロナワクチンについて

https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/vaccine_pfizer.html

無作為でない割付け：観察研究による平均因果効果の推定



このような観察研究で大規模な追跡調査を行うには、**交絡因子**の影響を取り除く必要がある。

観察研究におけるデータ

id	結果	処置	X1	X2	...	Xp
1	1	0	170	66	...	320
2	1	0	165	58	...	480
3	0	1	155	45	...	490
4	0	1	143	38	...	510
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$n - 1$	1	1	181	81	...	290
n	0	0	176	85	...	450

交絡を取り除くために、**多数の共変量**を統制する必要がある。

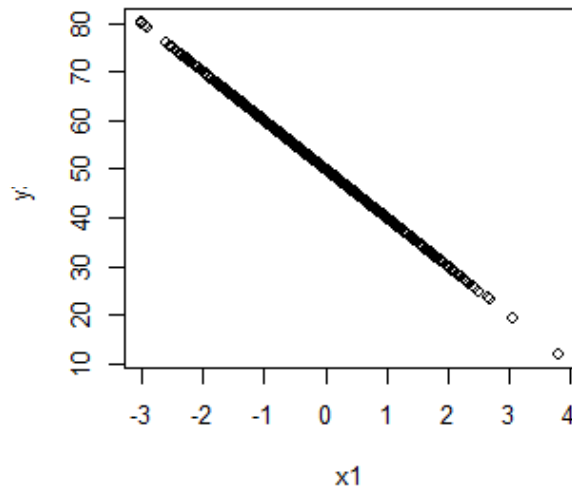
チョコレート消費量 と ノーベル賞受賞者数との関係

高橋 (2022, pp.77-89)

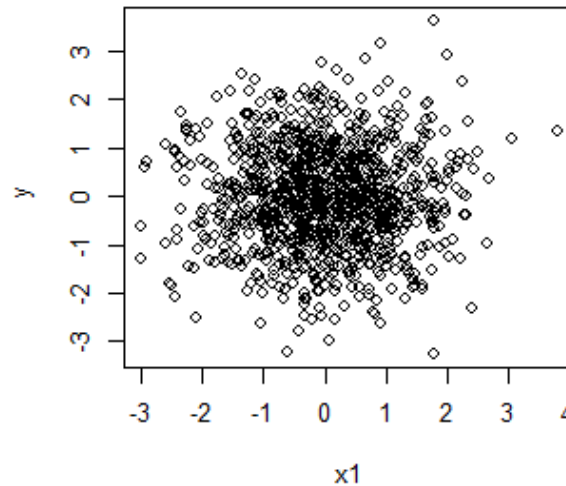
散布図と相関係数の例

$$-1 \leq r \leq 1$$

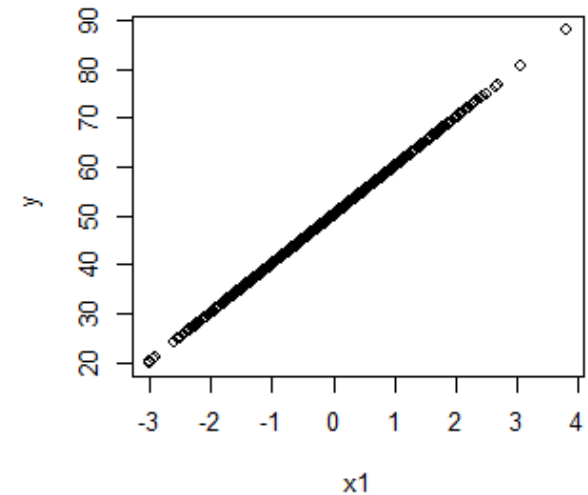
相関係数=-1.0



相関係数=0.0



相関係数=1.0

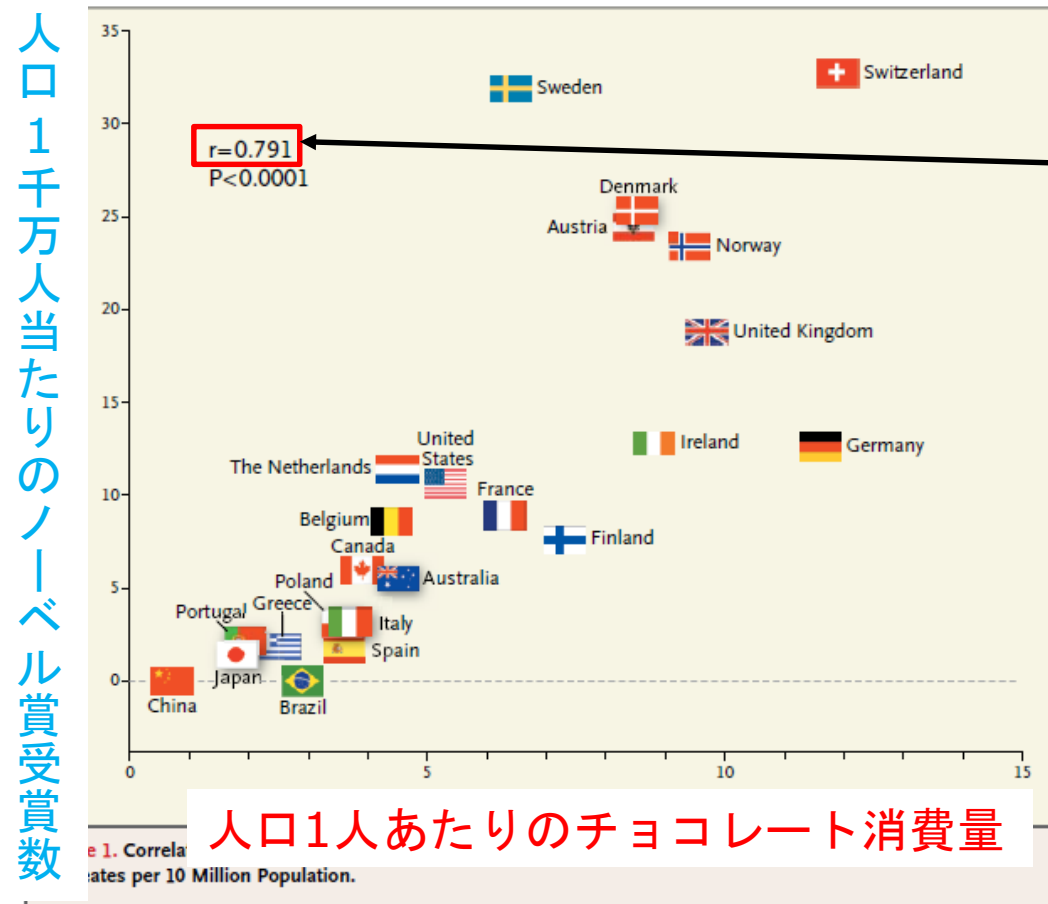


相関係数

2つの量的変数の線形的な関係性の強さを表す

散布図と相関係数

出典：Messerli, F. H. (2012) "Chocolate Consumption, Cognitive Function, and Nobel Laureates," The New England Journal of Medicine, 367 (16), pp.1562-1564.



$r = 0.791$

人口1人あたりのチョコレート消費量

単回帰モデル

```
> summary(modell<-lm(Nobel~Choco))
```

Call:

```
lm(formula = Nobel ~ Choco)
```

Residuals:

Min	1Q	Median	3Q	Max
-10.1603	-4.3915	-0.7202	2.5621	16.3355

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3.4217	3.2274	-1.060	0.301096
Choco	2.7044	0.5985	4.519	0.000188 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.65 on 21 degrees of freedom

Multiple R-squared: 0.493, Adjusted R-squared: 0.4689

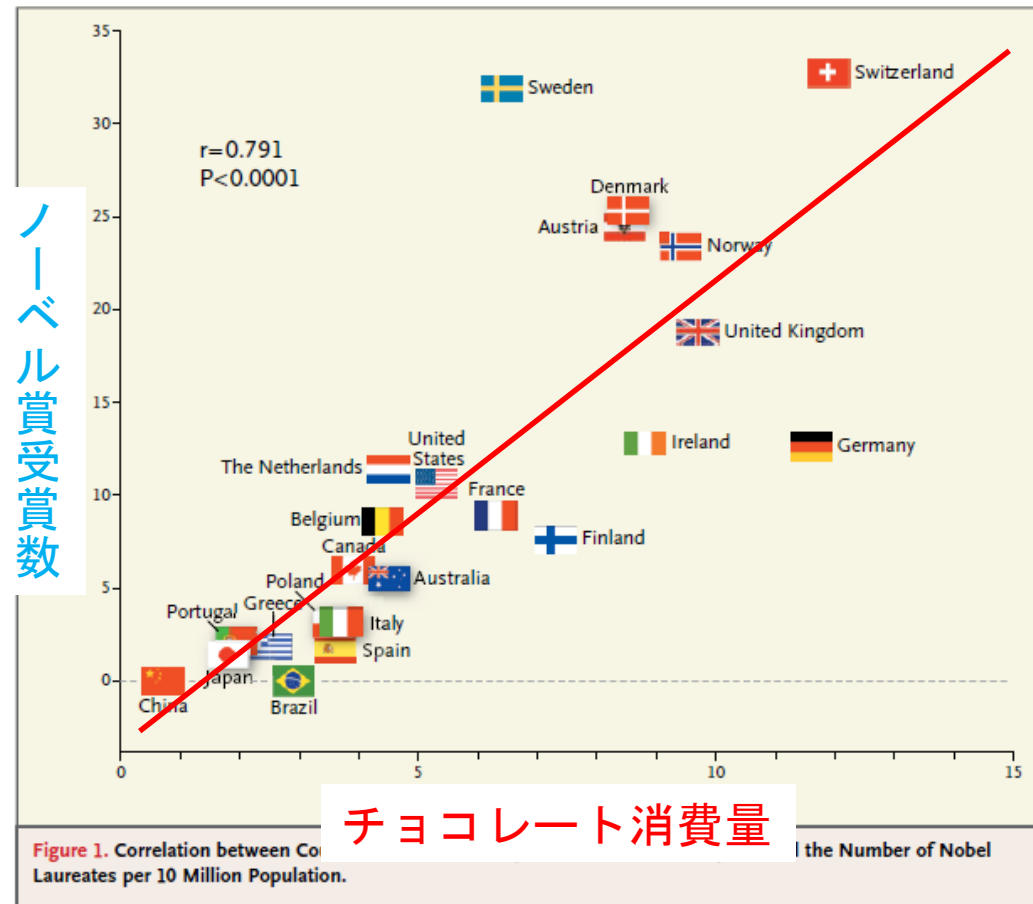
F-statistic: 20.42 on 1 and 21 DF, p-value: 0.000188

```
> confint(modell)
```

	2.5 %	97.5 %
(Intercept)	-10.133320	3.290018
Choco	1.459837	3.949020

相関

- 国ごとのノーベル賞受賞数とチョコレート消費量には正の相関がある.



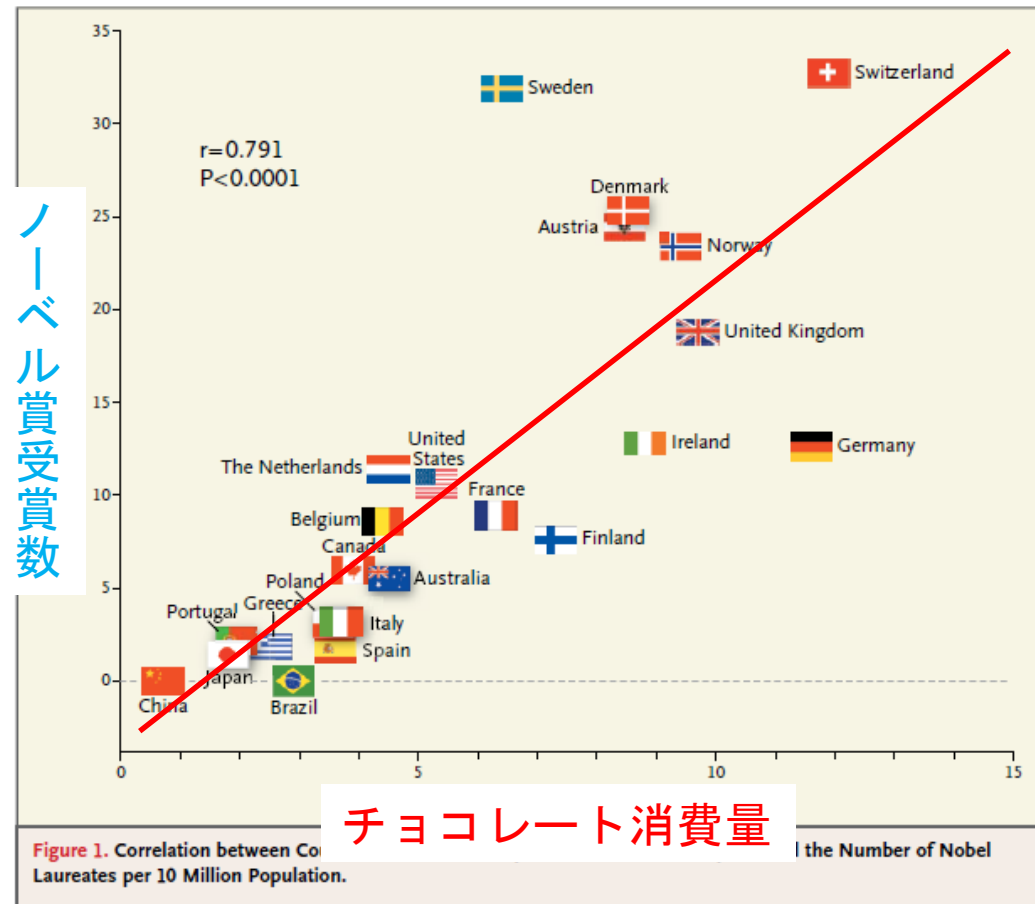
$$r = 0.791$$

チョコレートの消費量とノーベル賞受賞者数との関係

予測

$$\widehat{Nobel} = -3.422 + 2.704 \times Choco$$

- チョコレート消費量の多い国は、平均して、ノーベル賞受賞数が多いという傾向がある。



$r = 0.791$

未来を変えるには？

- 未来において日本のノーベル賞受賞数を増やすためには、今チョコレート消費量を増やせばよいのか？

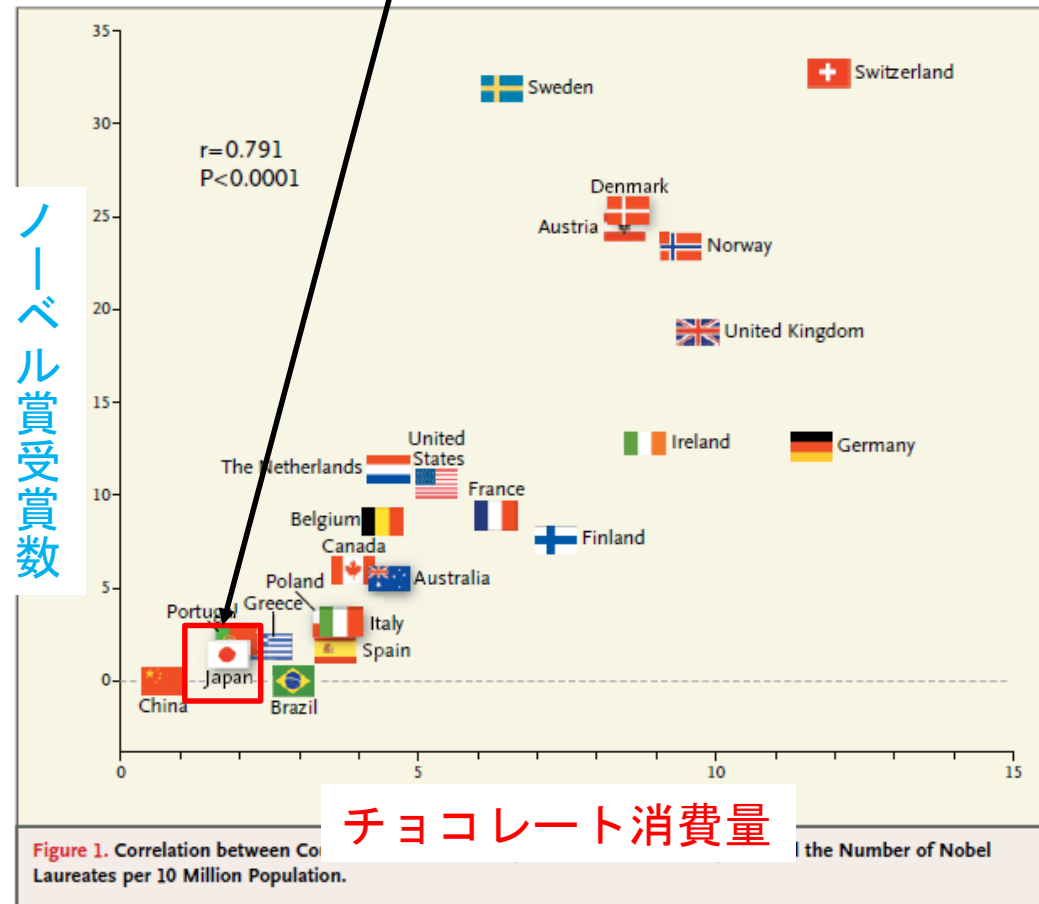
- 未来を変えるためには、今何をすればよいのか？
 - そのためには因果関係を明らかにしなければ答えが出ない。

相関係数に関する注意事項

- 相関関係と因果関係は必ずしも同じではない
- 因果関係（復習）
 - 要因Xを操作するとき、要因Yが変化すること
 - 操作なくして因果なし（no causation without manipulation）

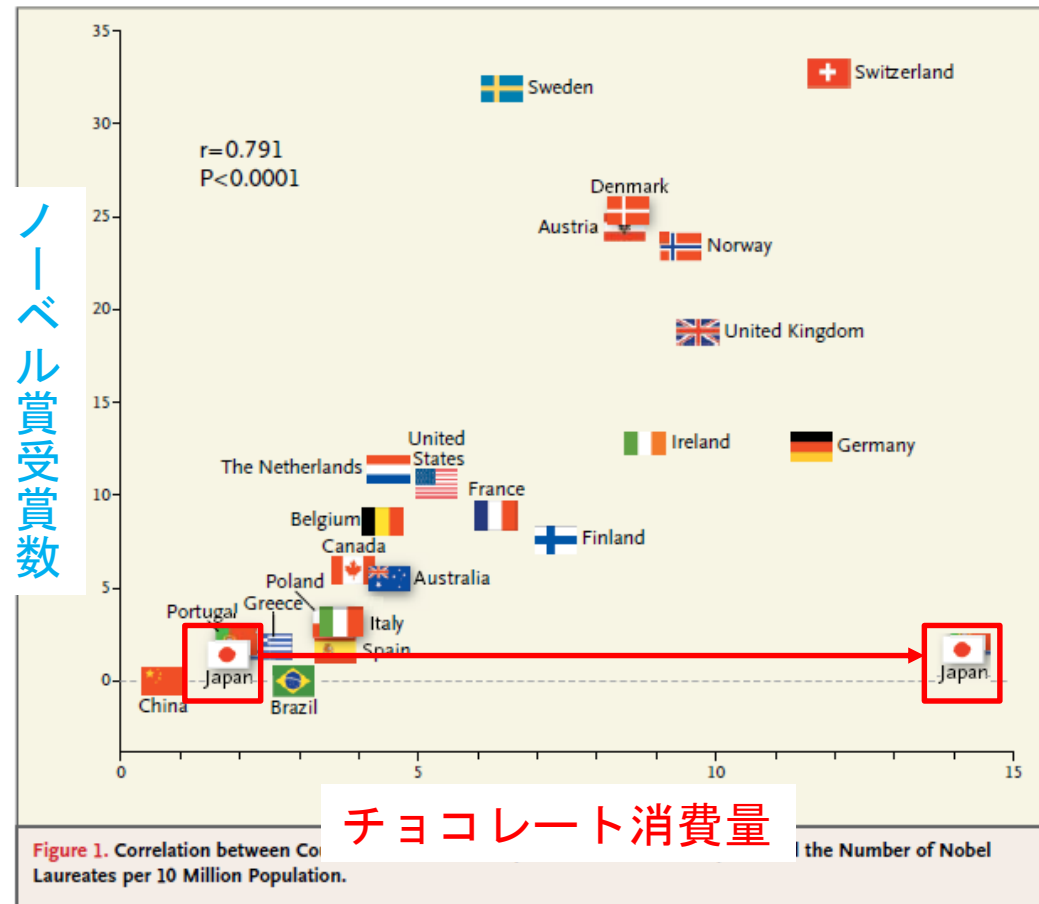
現状

日本の現状はここ



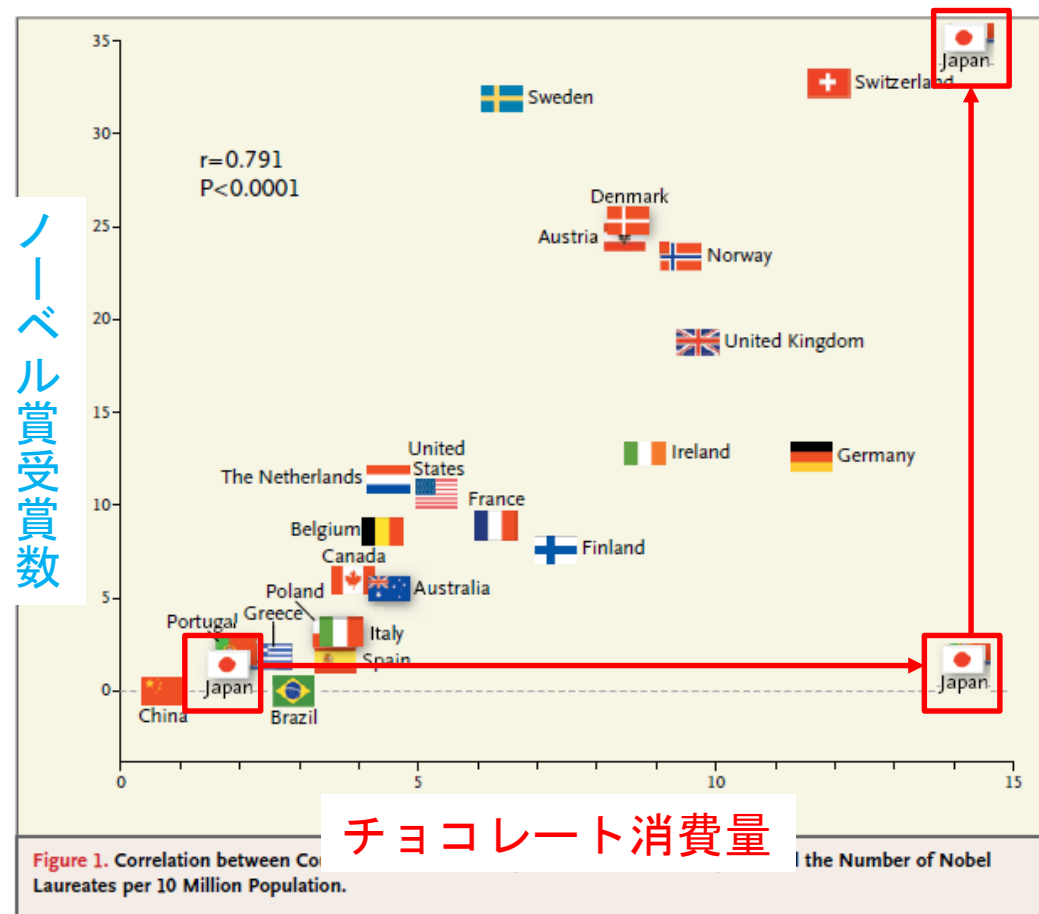
チョコレート消費量を操作

- 日本のチョコレート消費量を多くすると,



因果？

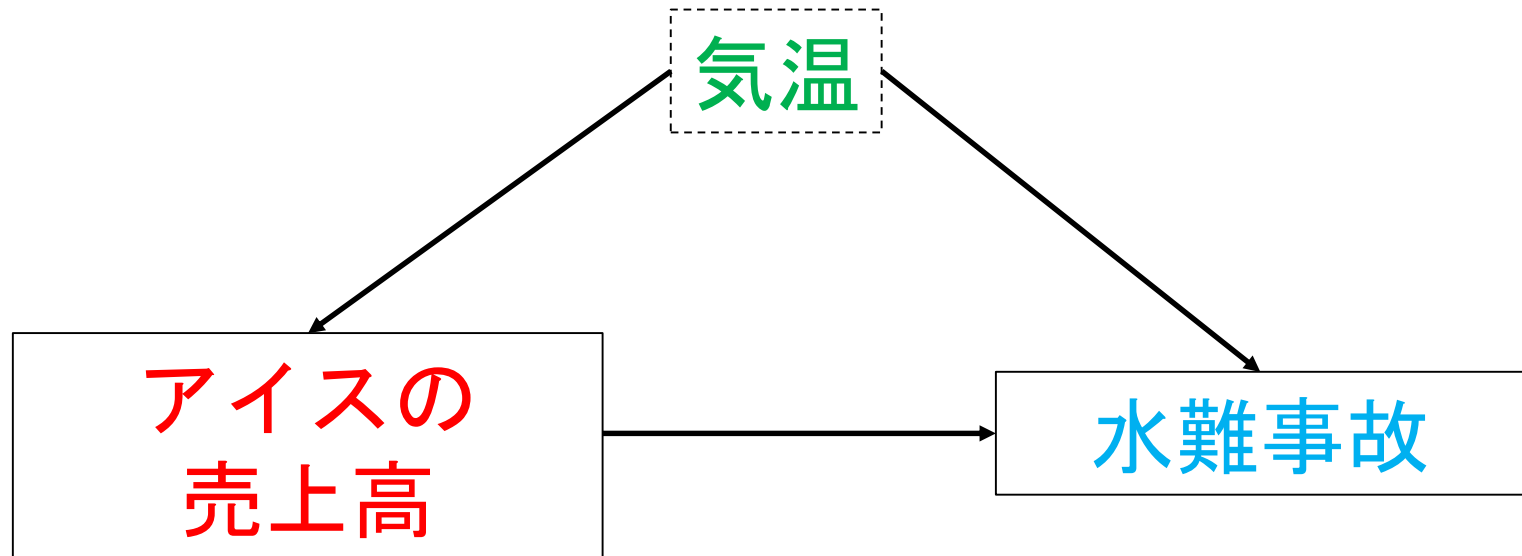
- 日本のチョコレート消費量を多くすると、日本のノーベル賞受賞数は増える？



?

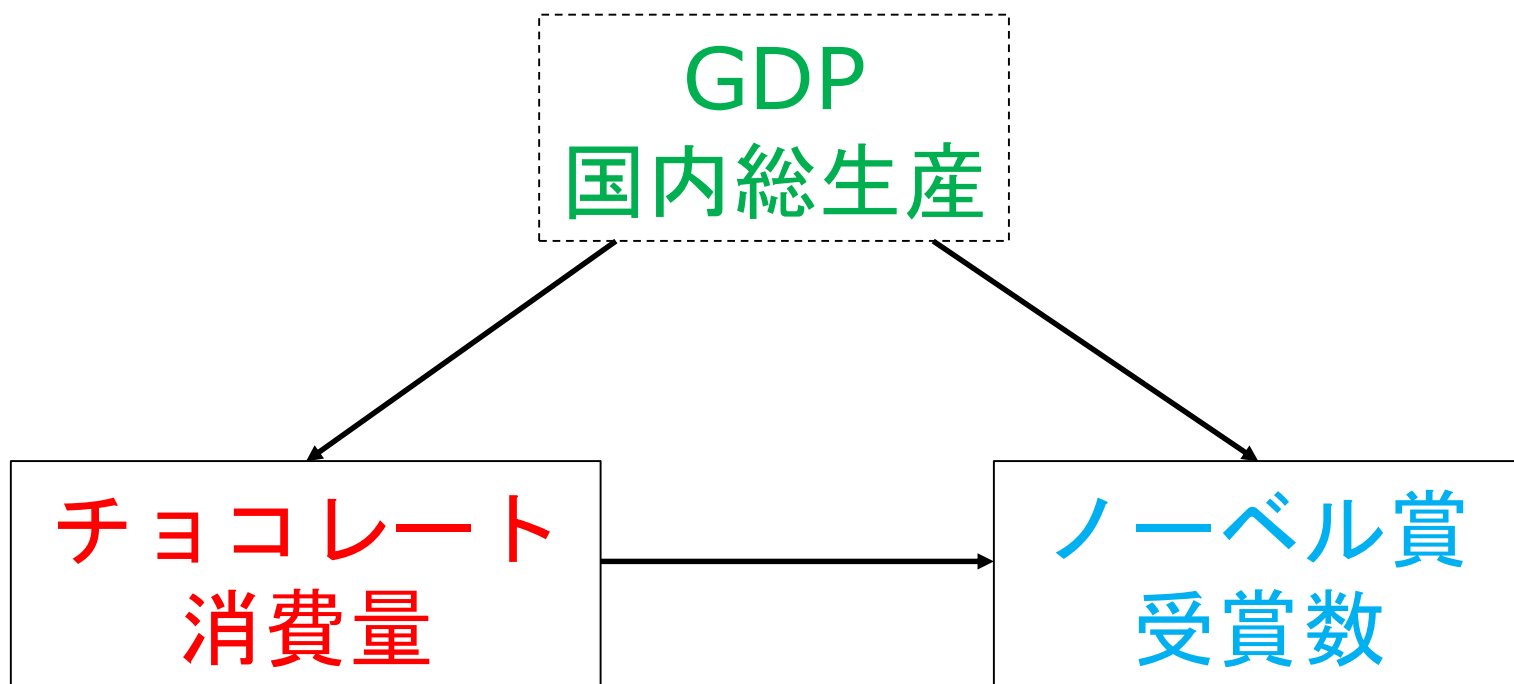
疑似相関

- 相関係数の絶対値が1に近いにもかかわらず、実際には2つの現象に直接的な関係がないこと



第3の変数

- ノーベル賞受賞数とチョコレート消費量の背後に共通の原因であるGDP（国内総生産）があると考えられる.



チョコレートの消費量とノーベル賞受賞者数との関係

データの例

国名	ノーベル賞	チョコ消費	GDP
スイス	28.80	8.8	85.135
ノルウェイ	20.37	5.8	74.986
イギリス	18.78	7.6	41.855
フランス	9.05	4.3	40.319
カナダ	4.17	4.0	46.550
⋮	⋮	⋮	⋮
日本	1.98	1.2	40.063
中国	0.02	0.1	10.004

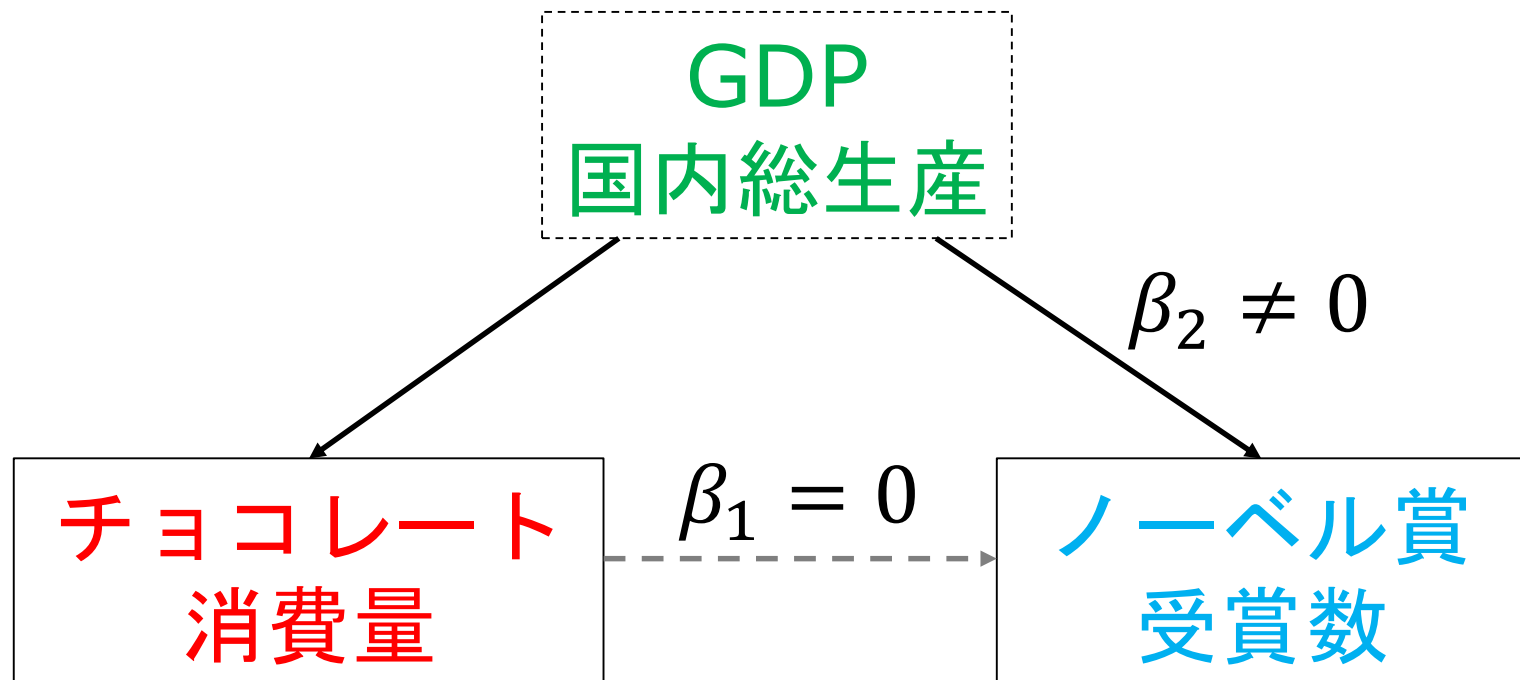
注：ノーベル賞は人口1千万人当たりの受賞者数，チョコ消費は人口1人当たりの消費量（kg），GDPは1人当たりの国内総生産（単位：1000米ドル）

重回帰モデル

$$\hat{Y}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

ノーベル賞受賞数_i

$$= \beta_0 + \beta_1 \text{チョコレート消費量}_i + \beta_2 \text{GDP}_i$$



用語の整理

□ $\hat{Y}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$

□ Y : 結果変数

■ 被説明変数, 目的変数, 従属変数, 応答変数

□ X_1 : 処置変数

□ X_2 : 共変量 (交絡因子)

■ 共変量だが交絡因子でない変数については, 講義4で解説する.

統計的因果推論の入り口

重回帰分析の限界

$$\square \hat{Y}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \underbrace{\beta_3 X_{3i} + \cdots + \beta_p X_{pi}}_{\text{いろいろな変数をモデルに入れる}}$$

いろいろな変数をモデルに入れる

仮定1：誤差項の期待値ゼロ

仮定2：パラメータ（母数）における線形性

仮定3：誤差項の条件付き期待値ゼロ

仮定4：完全な多重共線性がないこと

仮定5：誤差項の分散均一性

特に、仮定2と仮定3を満たさないと因果推論はできない！！

高橋（2022, pp.90-135）

傾向スコア (propensity score)

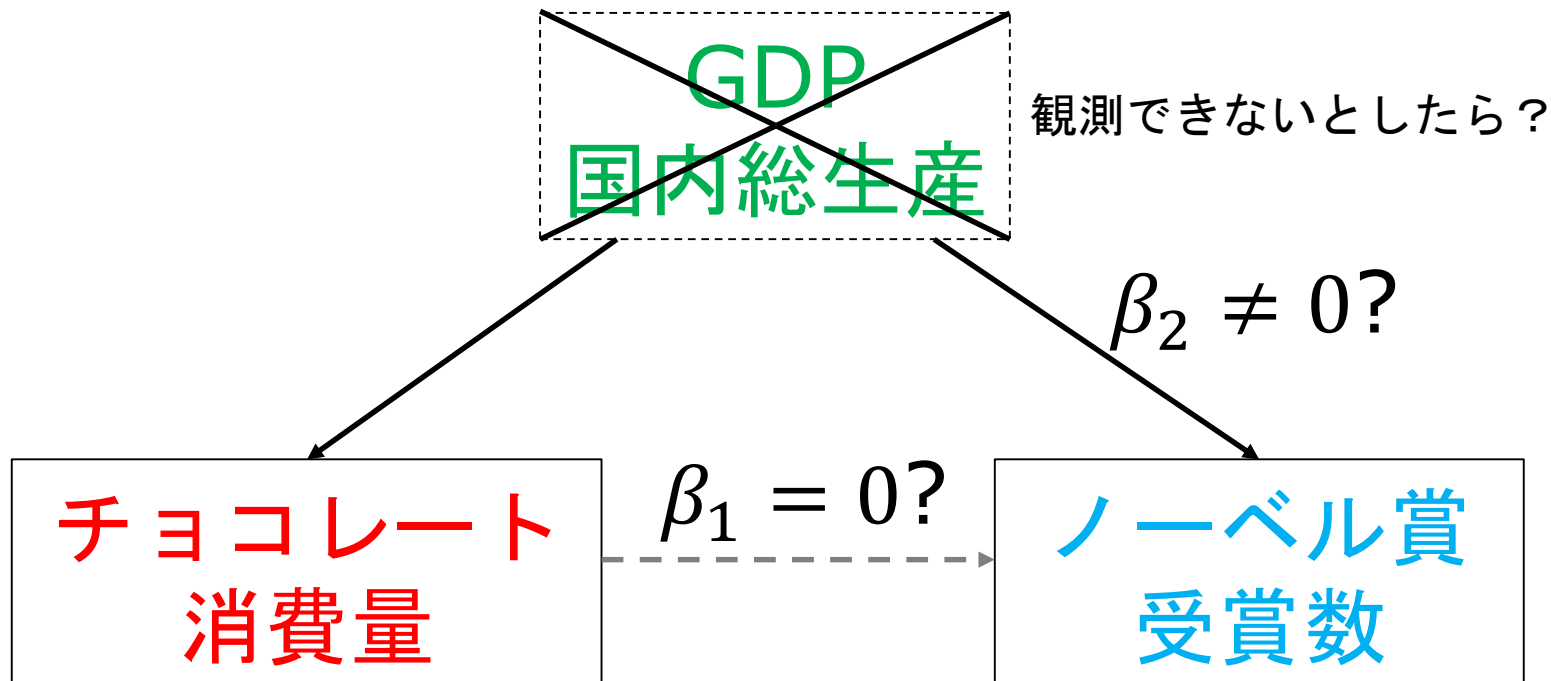
$$e(X) = \Pr(T_i = 1|X)$$

- 共変量 X が与えられたとき，処置に割付けられる確率
- 共変量 X を条件としたときに， $T_i = 1$ となる確率

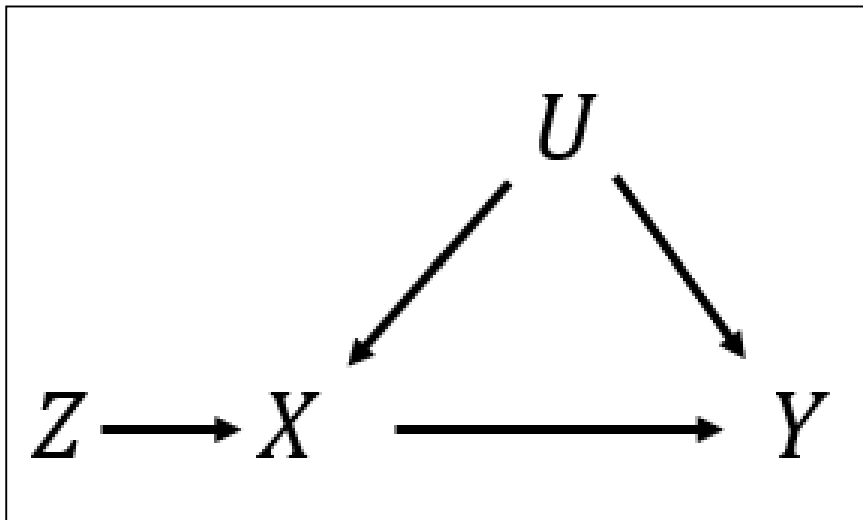
高橋 (2022, pp.136-182)

第3の変数が利用できないとき

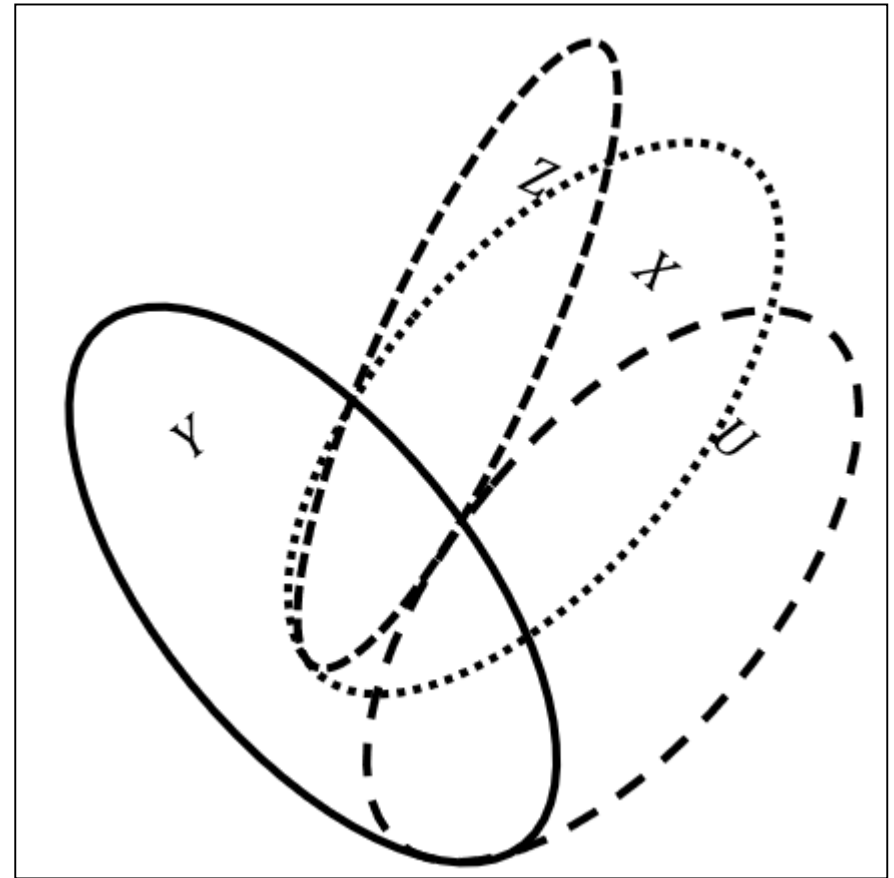
- 調整ができず第3の変数（潜在変数）の影響を考慮できない
- データ収集の段階で、重要な変数を見落とさないことが必要



操作変数法のイメージ図



操作変数 Z があれば、
未観測の交絡因子 U の
影響を統制できる。



傾向スコア的前提が満たされない例

□ 試験の点数

- 60点未満の学生は補習授業を受ける
- 60点以上の学生は補習授業を受けない

□ 処置の割付けは確定的

- 処置群と統制群の間に重なりがない
- このような研究課題は、傾向スコアを用いて適切に解析できない

回帰不連続デザイン

- RDD: regression discontinuity design
 - 回帰不連続デザイン
 - 回帰分断デザイン
 - 回帰非連続デザイン
 - 不連続回帰デザイン
- 回帰不連続デザインも準実験の1つ
 - 処置の割付けが確定的なときに使用できる
 - さまざまな準実験の中で最も実験研究に近いデザインといわれる
 - 局所的な範囲で無作為割付けが成立していると思なすことができるから