計算社会科学入門

水野貴之

E-mail: mizuno@nii.ac.jp

目次

はじめに

- 自然・社会の観察によって得られるデータは興味の対象以外の様々な影響を受けてデータが生成されるから，観察データから要因と結果の関係を知るには統計的な注意が必要
  - ニュースコンテンツを読んだ回数が多いほど政治知識が少ないという結果が得られた
  - ニュースコンテンツは政治知識を減らす効果があるのだろうか？
  - 「No」メディアのニュースコンテンツをあまり読まない人は新聞をよく読むという傾向があったからである.
- 社会の現象を扱う場合，人種/性差別を含む統計的差別・医療デマ・政策/政権の偏った評価・誤った経済対策などの，社会や人の意思決定に大きな影響を与え得る問題が発生しかねないため，十分な注意が必要

統計モデル

- 統計モデリングとは結果y（目的変数）と要因$\vec{x}$（説明変数）の関係をモデル化すること

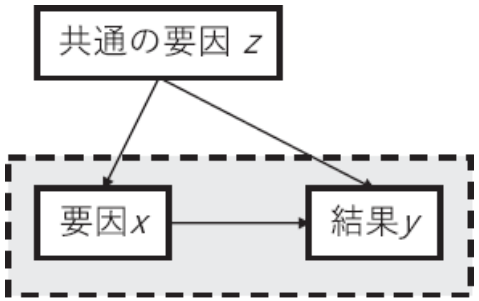
$$y \sim P(\vec{\theta})$$
$$\vec{\theta} = f(\vec{x}|\vec{\beta})$$

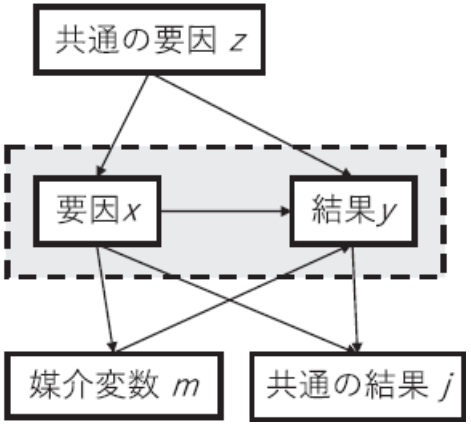
ここで$P(\vec{\theta})$は何らかの確率分布，$\vec{\theta}$はその分布のパラメータ，$f(\vec{x}|\vec{\beta})$ は説明変数$\vec{x}$と確率分布のパラメータの関係を表す関数fとそのパラメータ$\vec{\beta}$である

- 的変数y がユーザの政治知識，説明変数$\vec{x}$が自社メディアのニュースコンテンツと新聞の利用頻度になる（それぞれx_1, x_2）．データにモデルが適切に当てはまっていれば，$\vec{\beta} = \{\beta_1, \beta_2\}$ がx_1 とx_2 のy への貢献度合い（重み）
- $\vec{\beta}$ は目的やモデルの複雑さに応じて最尤推定法やマルコフ連鎖モンテカルロ法(MCMC) などを用いてアルゴリズムによって求める
  - 基本的には$f(\vec{x}|\vec{\beta})$ というモデルの下で，観測したデータ$\vec{x}$ から観測したデータy を最も高い確率で生成できるようにパラメータ$\vec{\beta}$ を求める．

興味のある要因変数以外の変数の影響

- 行動ビッグデータは観察データであるため興味のある変数間の関係以外にも考慮しなければならない変数が存在
- 考慮しなければいけない変数を見落とすと，効果を過大に評価したりする
- 図では，モデルに共通の要因z を組み込むことで，興味のある要因x の効果のより妥当な評価ができる
- モデルに組み込むべき変数はどのように選ぶべきか？
  - 対象のデータを生成する現象を考察して，そのメカニズムをモデル化する形で変数と変数間の関係（パス図）を描くのがよい
  - 目的変数と説明変数の関係や，説明変数間の関係を調べて候補を洗い出し，そのうえで現象のメカニズムについて仮説を立て，変数を選択していくことも有効





冪分布に従う目的変数

- 目的変数の大きな偏りも行動ログデータの特徴
  - 友人の数，Twitterのリツイート数，商品の販売数，資産，各友人とのコミュニケーション回数は冪分布と呼ばれる非常に偏った分布になる

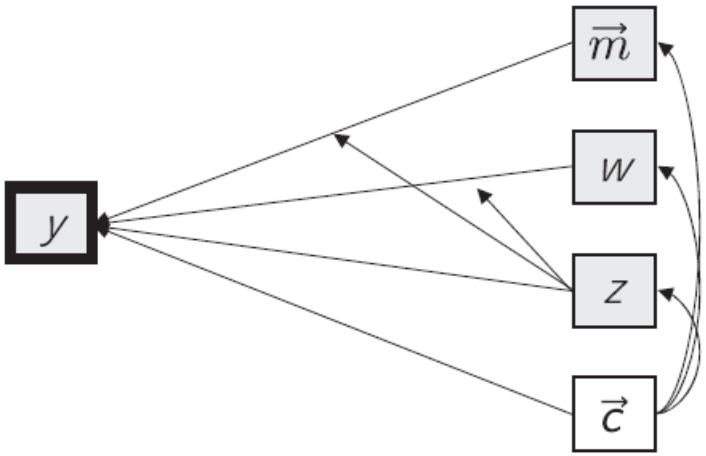
$$p(y) \propto y^{-\gamma}$$

  - $p(y)$の発生確率がyの冪乗に比例する．$2 \leq \gamma \leq 3$が多い
  - 冪分布に従う目的変数の扱いは厄介
    - 統計学の教科書で冪分布に言及されていることもほとんどなく，多くの統計ツールもモデルの目的変数の分布として冪分布を用意していない．一方で，最も拡散された情報や最も売れた商品を外れ値として無視するわけにもいかない．
- 生成過程
  - 富める者はますます富む(The rich get richer)
  - ランダムな乗算過程（自己フィードバック）が冪分布を生む

$$y_{t+1} = b_t y_t + \varepsilon_t$$

事例：ニュース画面への偶発的接触がニュース知識に与える影響

- インターネットテレビABEMA におけるチャンネル変更中の偶発的接触がニュース知識やほかのメディア利用効果に与える影響
  - 4 秒以下の視聴を偶発的接触
  - 偶発的接触頻度＝チャンネルを変えた回数
- は偶発的接触z とメディア利用頻度$\vec{m}$（マスメディア，オンラインニュース，まとめサイト，ソーシャルメディア）との交互作用がニュース知識y に与える影響を評価
- 短時間の偶発的接触がほかのメディアの利用によって強化される可能性を探る
- 「ニュースチャンネルではすぐチャンネルを変える」というニュースへの関心のなさを調整するためにニュースや政治への意識・態度に関わる変数を統制変数$\vec{c}$ として用いた
- 目的変数であるニュース知識はニュースに関するクイズの結果から項目反応理論によって求めた
  - ニュース知識yは平均0，標準偏差1 に正規化された連続値を取るため正規分布としてモデル化



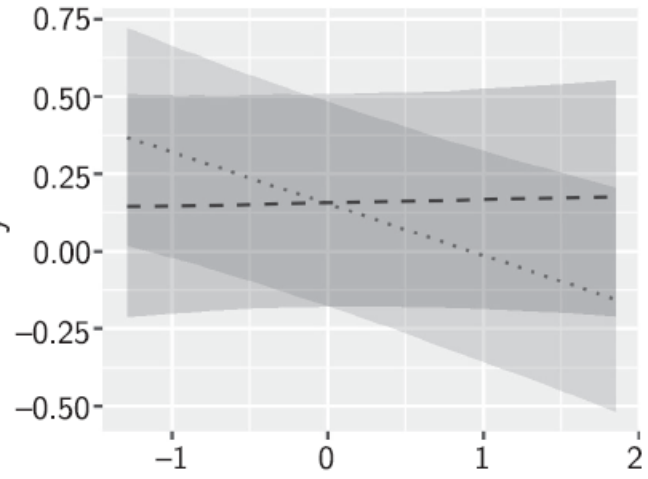
$$y \sim \mathrm{Normal}(\mu, \sigma)$$

$$\mu = \vec{\alpha} \cdot \vec{m} + \beta_1 \log_{10}(w+1) + \beta_2 \log_{10}(z+1)$$
$$+ \vec{\gamma} \cdot \vec{m} \log_{10}(z+1) + \gamma' \log_{10}(w+1) \log_{10}(z+1)$$
$$+ \zeta \cdot \vec{c} + \gamma_0$$

- $\vec{m}$ はメディア利用頻度因子，w はABEMA ニュース視聴時間，z はABEMA ニュースへの偶発的接触，$\vec{c}$ はそのほかの統制変数である
- w, z, w', z' は偏った分布を示したので，統計モデルで利用する際には対数変換して用いる
- 視聴時間や偶発的接触回数は0 の場合もあるので対数変換の際には$\log_{10}(x+1)$ のように1 を足している.
- $\vec{m} \log_{10}(z+1)$ はメディア利用頻度因子とニュースへの偶発的接触との交互作用
- $\log_{10}(w+1) \log_{10}(z+1)$ はニュース視聴時間とニュースへの偶発的接触回数との交互作用

- 偶発的接触(incidental exposure) の主効果には明確な傾向は見られなかった
- 偶発的接触とメディア利用頻度の交互作用においては，偶発的接触とソーシャルメディアのみ明確な正の効果が見られた．この交互作用を直観的に可視化するために，偶発的接触が少ない(25 %ile) 視聴者と多い(75 %ile) 視聴者がソーシャルメディアの利用頻度を変えたときの，統計モデルによるニュース知識y の予測値を示す

カテゴリ	説明変数	中央値	2.5 %ile	97.5 %ile
ABEMA News	視聴時間 $\log_{10} w$	0.001	[−0.067,	0.068]
	偶発的接触 $\log_{10} z$	0.003	[−0.080,	0.083]
メディア利用 $\vec{m}$	ソーシャルメディア	**−0.075**	[−0.141,	−0.010]
	まとめサイト	−0.003	[−0.074,	0.070]
	マスメディア	**0.108**	[0.037,	0.180]
	オンラインニュース	0.026	[−0.058,	0.109]
$\log_{10} z$ と $\vec{m}$ の交互作用	ソーシャルメディア	**0.111**	[0.050,	0.171]
	まとめサイト	−0.012	[−0.086,	0.059]
	マスメディア	0.006	[−0.058,	0.071]
	オンラインニュース	−0.035	[−0.121,	0.049]
$\log_{10} z$ と $\log_{10} w$ の交互作用	ABEMA News 視聴時間	0.003	[−0.057,	0.063]



モデルの比較

- 行動データでは様々な項目のデータが取られていることが多いため複雑なモデルも容易に作ることができる.
- 複雑すぎるモデルは柔軟性が高く，いくらでもデータに合わせることができてしまうため，不適切な分析結果をもたらす.
  - 説明変数が多いい）ほど調整の余地が大きくなるため，当てはまりはよくなる.
    - 例えば目的変数とまったく関係のない変数を説明変数（例えば乱数）として加えるだけでも，偶然発生するゆらぎと目的変数の相関によって当てはまりはよくなる
- 適切に説明変数を選ぶ際に有用なのが情報量基準によるモデル選択
  - 代表的な情報量基準として赤池情報量基準(AIC: Akaike Information Criteria)

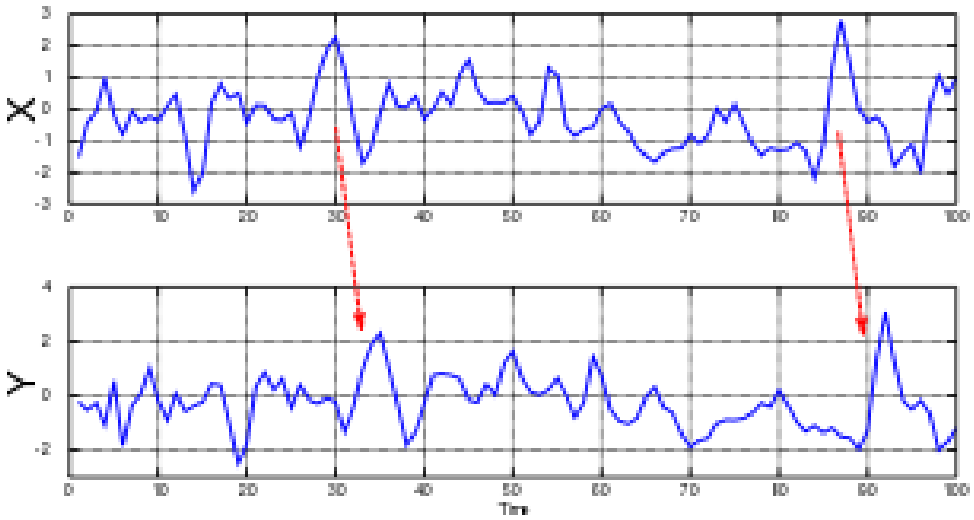
$$AIC = -2\log L + 2k$$

L はモデルの当てはまりの指標（尤度），k は回帰分析や一般化線形モデルでは説明変数の数

  - 当てはまりがよいほど（L が大きいほど）AIC はそのモデルをよいモデル
  - 説明変数が少ないほど（k が小さいほど）AIC はそのモデルをよいモデル

グレンジャー因果性

- グレンジャー因果検定は、ある時系列が別の時系列の予測に役立つかどうかを判断するための統計的仮説検定で、1969年に初めて提案された。
- 「因果関係」という用語を単独で使用することは誤用で、グレンジャー因果性は「前後関係（precedence）」、またはグレンジャー自身が1977年に主張したように「時間的な関連（temporally related）」と説明される方が適切である。グレンジャー因果性は、XがYを引き起こすかどうかを検定するのではなく、XによってYを予測できるかどうかを検定するものである
- 「因果⇒グレンジャー因果」であることに注意！



時系列変数Xから時系列変数Yへのグレンジャー因果がある

グレンジャー因果検定の手順

1. 時系列変数Xの定常性検定（ADF検定）
2. 時系列変数Yの定常性検定（ADF検定）
3. 時系列変数Xと時系列変数Yとで単位根を持つ帰無仮説が、
   - 両方とも棄却される→生の値でGranger causality検定へ
   - 片方のみ棄却される→両方とも1階差をとりGranger causality検定へ
   - 両方とも棄却されない→共和分検定へ
4. 時系列変数Xと時系列変数Yとの共和分検定
5. 時系列変数Xと時系列変数Yとに共和分関係がない帰無仮説が
   - 棄却される→Error correction model（共和分成分を入れたGranger causality検定）へ

$$\Delta y_t = \mu + \theta_1 \left(y_{t-1} + \frac{\theta_2}{\theta_1} x_{t-1} \right) + \sum_{i=1}^{n} \alpha_i \Delta y_{t-i} + \sum_{i=0}^{m} \beta_i \Delta x_{t-i} + \mu_t$$

   - 棄却されない→両方とも1階差をとりGranger causality検定へ

ここで、1階差とは差分ΔY(t) = Y(t) − Y(t−1). 上記の1から5におけるパラメータ数の決定にはAIC を用いる.

定常過程

弱定常(単に定常)

条件:nが十分に大きいとき、平均値、分散、共分散が時間に関して不変である。

$$\text{平均値:}\quad \frac{1}{n-1}\sum_{t=0}^{n} x_t = \frac{1}{n-1}\sum_{t=T}^{n+T} x_t = \text{一定}$$

$$\text{分散:}\quad \frac{1}{n-1}\sum_{t=0}^{n} x_t^2 = \frac{1}{n-1}\sum_{t=T}^{n+T} x_t^2 = \text{一定}$$

$$\text{共分散:}\quad \frac{1}{n-1}\sum_{t=0}^{n} \left(x_t \cdot x_{t+\tau}\right) = \frac{1}{n-1}\sum_{t=T}^{n+T} \left(x_t \cdot x_{t+\tau}\right) = \text{一定}$$

データのどの部分で平均や分散，共分散を計算しても同じ値である。

強定常

x_t^3から求められる**歪度**，x_t^4から求められる**尖度**，・・・，と全てのモーメントが時間に関して不変であることはもちろんのこと，全ての変数間の関係も時間に関して不変である

同時分布関数が時間に対して不変

$$f(x_i, x_{i+1}, \cdots, x_n) = f(x_{i+T}, x_{i+1+T}, \cdots, x_{n+T})$$

AR(P)過程

統計学において時系列データに適用されるモデルの1つである.

AR(p)過程という表記は次数pの自己回帰モデルを表す.

$$X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + c + \varepsilon_t$$

ここで、φ_iとcは定数、ε_tは平均ゼロの確率変数(iidノイズ)

定常な時系列にAR(p)モデルを当てはめ，定数φ_iとcを求める方法は後ほど解説する.

ここでは，AR(1)過程を利用して定常性の検定をする，単位根検定について紹介する.

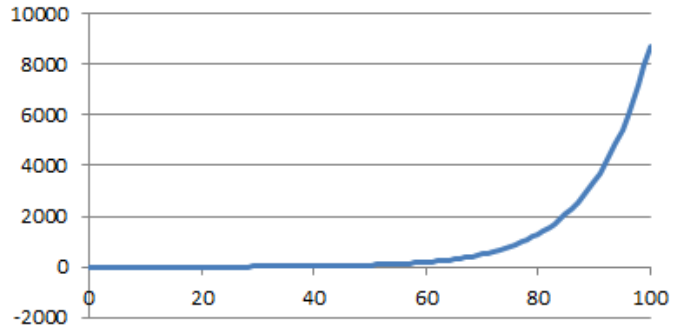
AR(1)過程

X_tが過去1つのX_{t-1}に依存ししている確率過程

$$X_t = \varphi_1 X_{t-1} + c + \varepsilon_t$$

ここで、φ_1とcは定数、ε_tは平均ゼロの確率変数（iidノイズ）

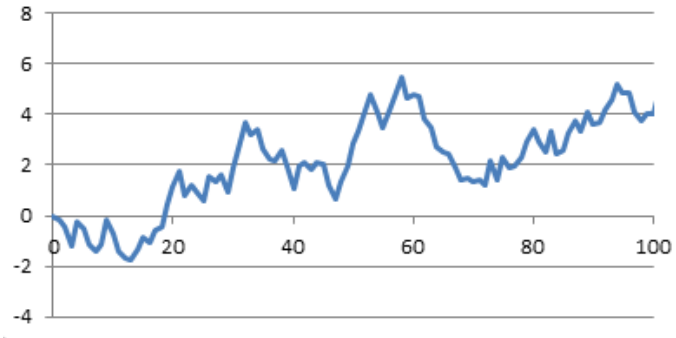
$X_t = 1.1X_{t-1} + \varepsilon_t$



$\varphi_1 > 0$の場合、指数関数的にXが大きくなる。
明らかにXの平均値が、時刻に依存する。

⇛ 非定常時系列

$X_t = X_{t-1} + \varepsilon_t$

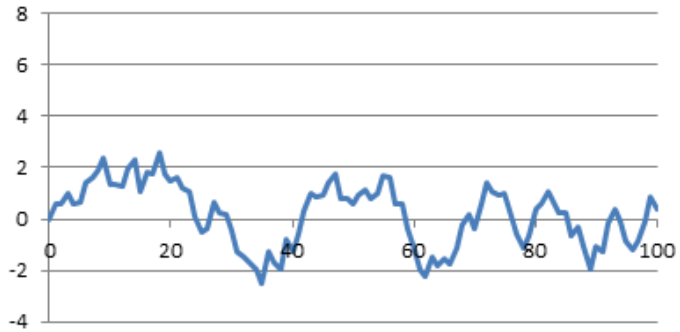


$\varphi_1 = 0$の場合、Xはランダムウォークする。
Xの平均値が、時刻に依存する。

⇛ 非定常時系列

$\varphi_1 < 1$場合

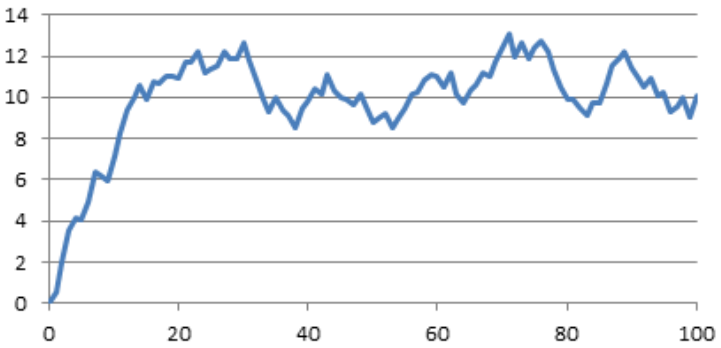
$$X_t = 0.9X_{t-1} + \varepsilon_t$$



$\varphi_1 < 0$ の場合、Xはある値のまわりで変動する。
Xの平均値が収束する。

⇛ 定常時系列

$$X_t = 0.9X_{t-1} + 1 + \varepsilon_t$$

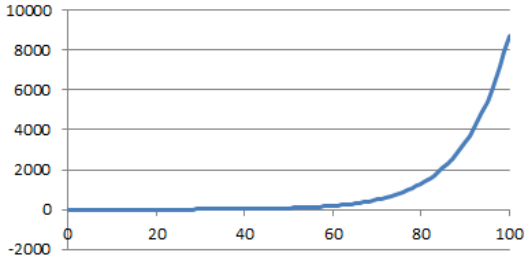


$c \neq 0, \varphi_1 < 0$の場合も、Xはある値のまわりで変動する。
Xの平均値が収束する。

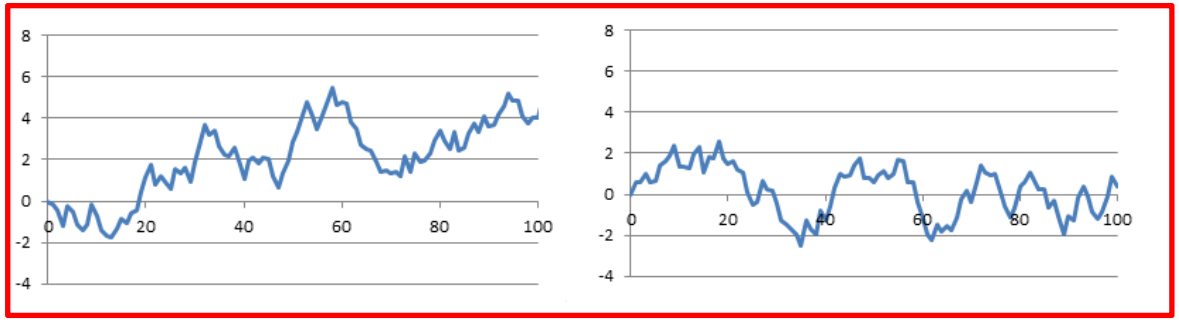
⇛ 定常時系列

単純な単位根検定(検出精度が悪い)

時系列の定常性を検定する方法



明らかに非定常



区別が難しい(左は非定常、右が定常)

AR(1)過程を用いた単位根検定では、上のような時系列が与えられた場合、時系列にAR(1)過程を仮定し、AR(1)過程を時系列に当てはめ、最小2乗法を用いて

$$X_t = \varphi_1 X_{t-1} + c + \varepsilon_t$$

パラメータφ_1を推定し、φ_1=1である帰無仮説を下記のt値を使って検定する。

$$t = \frac{\text{推定された}\varphi_1 - 1}{\text{推定されさ}_{\varphi_1}\text{の標準誤差}}$$

もしφ_1=1 であれば、このt値がt分布で近似できることを用いて検定する。

自由度とT分布表

自由度＝(データのサンプル数)－(パラメータ数)
＝100－2＝98

自由度	有意水準5%	有意水準1%
1	12.706	63.657
2	4.303	9.925
3	3.182	5.841
5	2.571	4.032
10	2.226	3.169
20	2.086	2.845
30	2.042	2.750
60	2.000	2.660
98	1.986	2.629
120	1.980	2.617
∞	1.960	2.576

もし帰無仮説が成立するのであれば、得られたt値は95%の確率で、

$$-1.986 \leq t\text{値} \leq 1.986$$

の範囲に入り、99%の確率で、

$$-2.629 \leq t\text{値} \leq 2.629$$

に入る。従って、これらの範囲外にあれば、帰無仮説は有効水準？%で棄却され、対立仮説「$\varphi_1 \neq 1$」が採択される。

t値が有意水準よりも大きければ、時系列が定常であると考えることができるが、
実は、、、

「~~もし$\varphi_1=1$であれば、このt値がt分布で近似できる~~」

あまりよい近似ではない。

デッキー・フェラー検定(DF検定)

$$X_t = \varphi_1 X_{t-1} + c + \varepsilon_t$$

を式変形して、

$$X_t - X_{t-1} = (\varphi_1 - 1)X_{t-1} + c + \varepsilon_t$$
$$\Delta X_t = \rho X_{t-1} + c + \varepsilon_t$$

ρ は先程よりt分布に近くなる。時系列に上式を当てはめ、最小2乗法を用いて、パラメータρ を推定し、ρ=0 である帰無仮説をt値を使って検定する。

$$t = \frac{\text{推定された}\rho}{\text{推定された}\rho\text{の標準誤差}}$$

このt値が有意水準？%よりも大きければ、帰無仮説が有意水準？%で棄却され、時系列が定常であると考えることができる。

「~~時系列にAR(1)過程を仮定し~~」

時系列が、AR(1)過程に従うとは限らいない。そこで、、、

ADF検定（拡張されたデッキー・フェラー検定）

ランダムウォーク	$\Delta X_t = \rho X_{t-1} + \varepsilon_t$
ドリフト付きランダムウォーク	$\Delta X_t = \rho X_{t-1} + c + \varepsilon_t$
ドリフト＋トレンド項	$\Delta X_t = \rho X_{t-1} + c + \beta t + \varepsilon_t$

DF検定では、通常、上記3つのAR(1)過程について$\rho = 0$であるかを検定し、$\rho \neq 0$でないことを統計的に示す。

ADF検定では、AR(2)、AR(3)など自己回帰係数のラグを増やした場合も含めて検定できるようにする。

例： AR(3)の場合

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + a_3 x_{t-3} + f_t$$

行列表示すると、

$$X_t = A X_{t-1} + F_t$$

$$X_t = \begin{pmatrix} x_t \\ x_{t-1} \\ x_{t-2} \end{pmatrix}, A = \begin{pmatrix} a_1 & a_2 & a_3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, F = \begin{pmatrix} f_t \\ 0 \\ 0 \end{pmatrix}$$

時系列が定常であれば、$A=\begin{pmatrix} a_1 & a_2 & a_3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$の固有値λが1より小さい。

固有値λが1であるかを帰無仮説として検定する。行列Aの固有方程式は、

$$\begin{vmatrix} a_1-\lambda & a_2 & a_3 \\ 1 & -\lambda & 0 \\ 0 & 1 & -\lambda \end{vmatrix} = \lambda^3 - a_1\lambda^2 - a_2\lambda - a_3 = 0$$

であるので、λ=1を代入して、

$$1 - a_1 - a_2 - a_3 = 0$$

であるかをt値を使って検定すればよい。以下のように式変形すると

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + a_3 x_{t-3} + f_t$$
$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + a_3 x_{t-2} - a_3 \Delta x_{t-2} + f_t$$
$$\Delta x_t = (a_1 + a_2 + a_3 - 1) x_{t-1} - (a_2 + a_3)\Delta x_{t-1} - a_3 \Delta x_{t-2} + f_t$$

となる。つまり、第1項の係数がゼロであるかをt値で検定することと同じ。

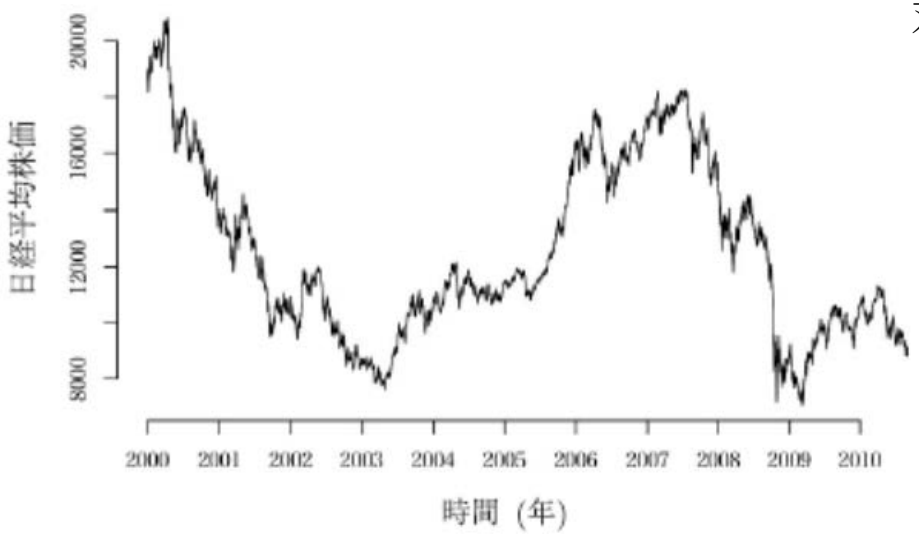
AR(p)では、

$$\Delta x_t = \rho_0 x_{t-1} + \rho_1 \Delta x_{t-1} + \rho_2 \Delta x_{t-2} + \cdots + \rho_{p-1} \Delta x_{t-p+1} + f_t$$

のρを推定し、$\rho_0 = 0$かの仮説検定を行う。

一般に、ランダムウォーク、ドリフト付き、ドリフト＋トレンドに関して、AR(1)～AR(4)まで（計12式を）調べて、定常性をチェックする。

株価データと弱定常性



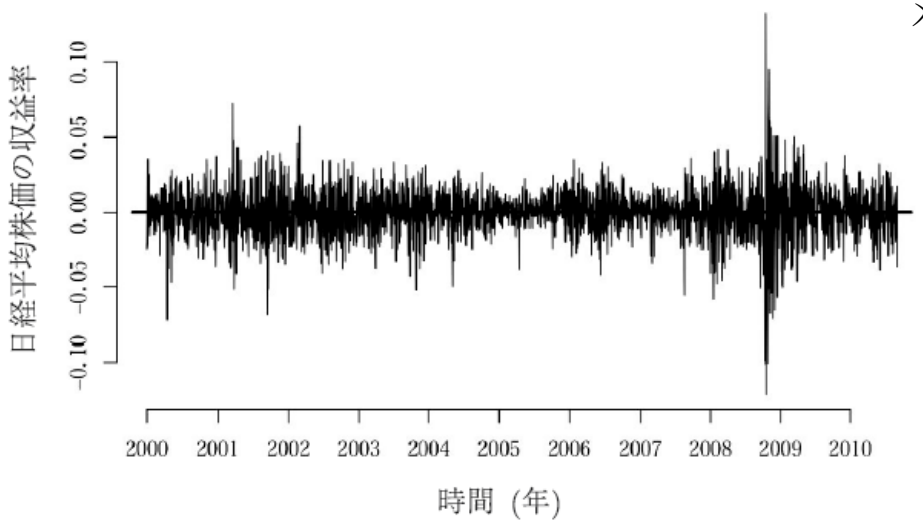
対数株価$\log S_n$

株価の動きに大きな方向性が存在

ADF検定のp値は0.5917
PP検定のp値は0.6124

どちらも5%有意水準は大きく超える

非定常



対数収益率$\log S_n - \log S_{n-1}$

大きな方向性は存在していない

ADF検定とPP検定のp値は、それぞれ0.01以下

どちらも5%有意水準は大きく超える

収益率は弱定常

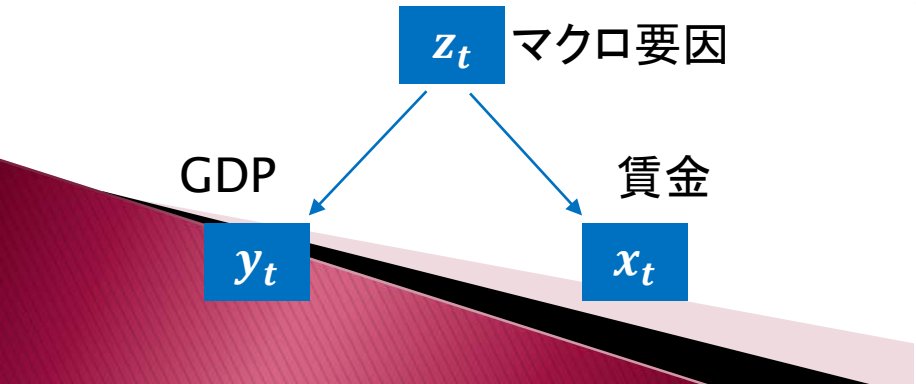
Cointegration検定

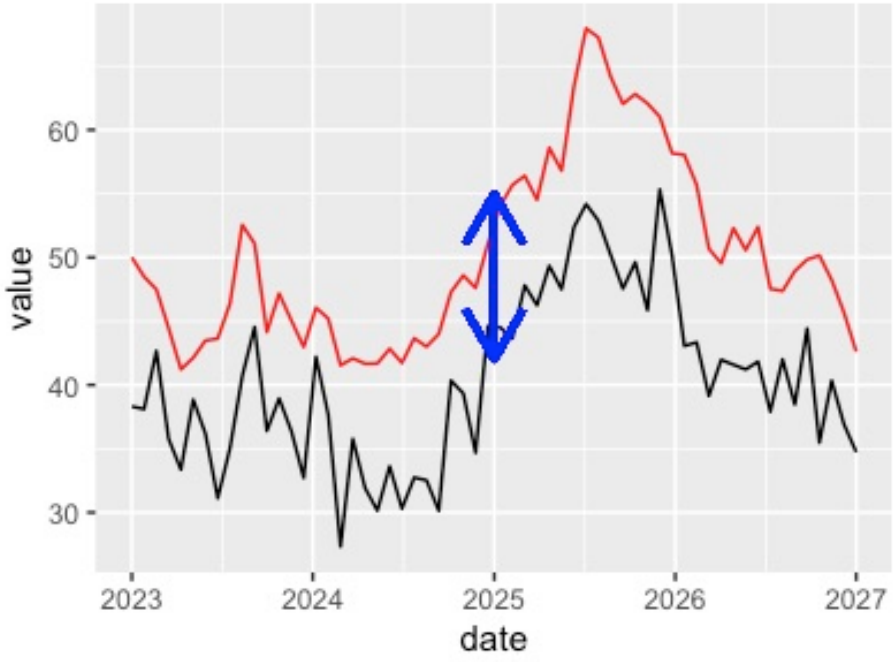
x_tとy_tが共和分（同期）しているならば、それらの変数のある線形結合は定常でなくてはならない。言い換えると、

$$y_t - \beta x_t = u_t$$

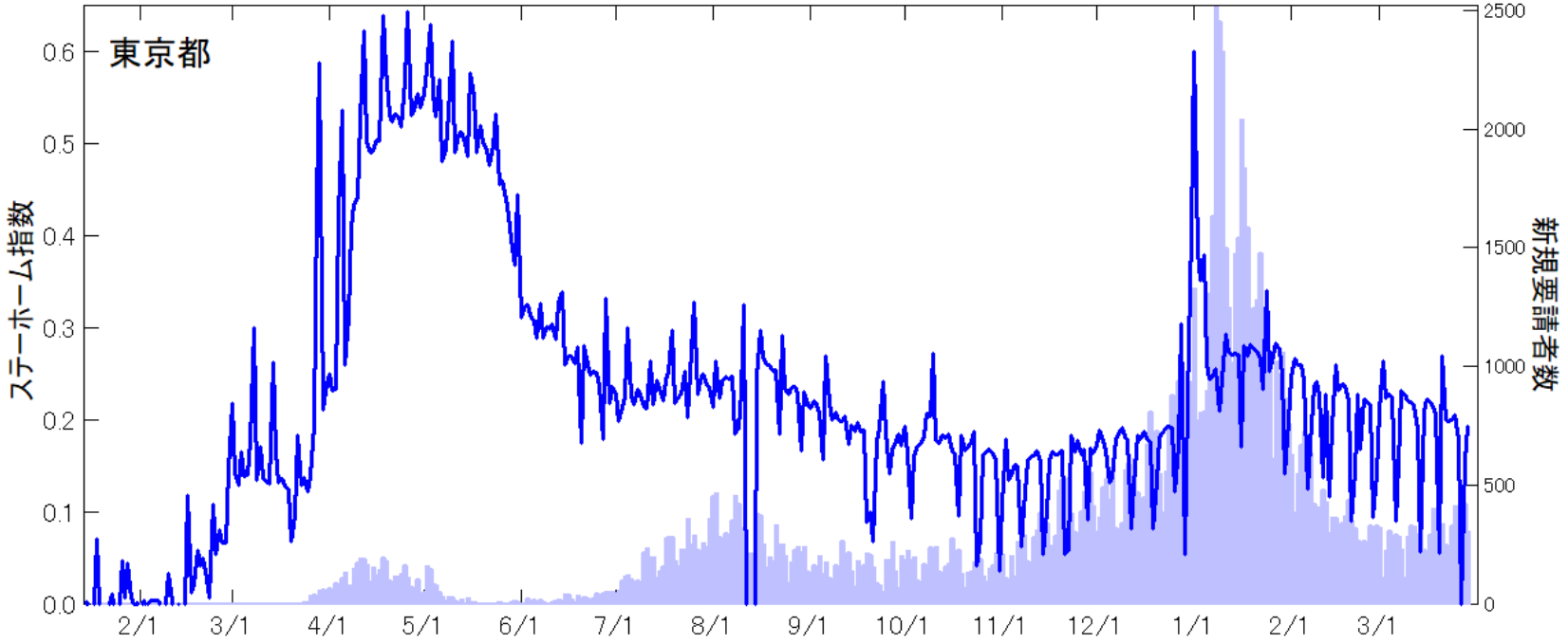
であり、ここでu_tは定常である。もし、u_tが分かっているのならば、ADF検定をおこなう。しかし、u_tは事前にはわからないのでまずそれを、一般的にはOLSを使って、推定する。そして推定したu_tに対して定常性の検定を行う。

アメリカの多数のマクロ経済時系列
（例えば、GNP、賃金、雇用者数など）
は共和分の関係があり、同期する

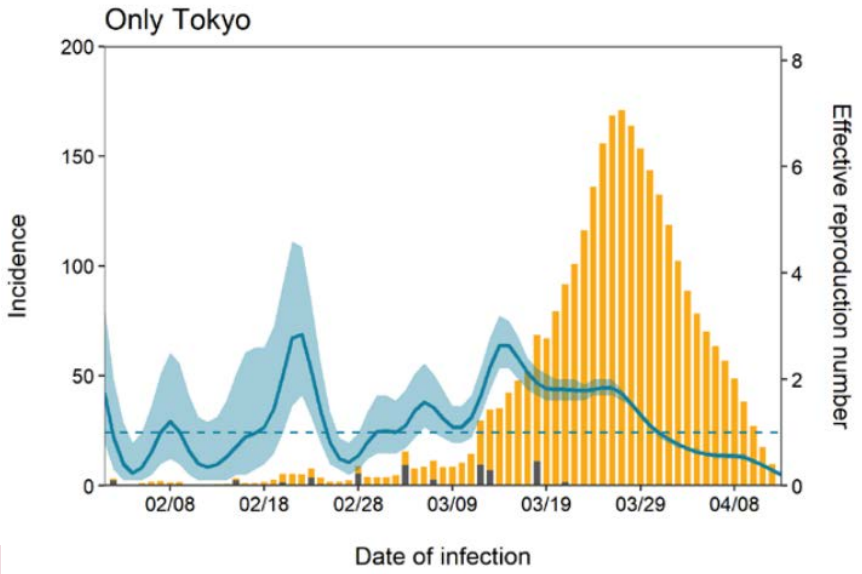




実効再生産数と外出の自粛率



SHR: Stay Home Rate
ERN: Effective Reproduction Number)



Prefecture	ADF of ERN	ADF of SHR	Coint
Hokkaido	0.000**	0.263	-
Aomori	0.015*	0.049*	-
Iwate	0.005**	0.016*	-
Miyagi	0.003**	0.073	-
Akita	0.000**	0.049*	-
Yamagata	0.078	0.089	0.216
Fukushima	0.003**	0.066	-
Ibaraki	0.002**	0.146	-
Tochigi	0.003**	0.315	-
Gunma	0.005**	0.240	-
Saitama	0.003**	0.465	-
Chiba	0.008**	0.516	-
Tokyo	0.004**	0.394	-
Kanagawa	0.003**	0.421	-
Niigata	0.000**	0.045*	-
Toyama	0.001**	0.114	-
Ishikawa	0.000**	0.164	-
Fukui	0.000**	0.115	-
Yamanashi	0.001**	0.259	-
Nagano	0.003**	0.207	-

Prefecture	ADF of ERN	ADF of SHR	Coint
Gifu	0.001**	0.152	-
Shizuoka	0.040*	0.110	-
Aichi	0.000**	0.344	-
Mie	0.000**	0.020*	-
Shiga	0.001**	0.156	-
Kyoto	0.014*	0.169	-
Osaka	0.013*	0.371	-
Hyogo	0.005**	0.225	-
Nara	0.000**	0.290	-
Wakayama	0.002**	0.092	-
Tottori	0.003**	0.000**	-
Shimane	0.038*	0.031*	-
Okayama	0.000**	0.061	-
Hiroshma	0.004**	0.163	-
Yamaguchi	0.000**	0.116	-
Tokushima	0.000**	0.003**	-
Kagawa	0.000**	0.034*	-
Ehime	0.001**	0.071	-
Kochi	0.000**	0.094	-
Fukuoka	0.000**	0.243	-
Saga	0.000**	0.068	-
Nagasaki	0.000**	0.023*	-
Kumamoto	0.000**	0.093	-
Oita	0.001**	0.015*	-
Miyazaki	0.002**	0.009**	-
Kagoshima	0.000**	0.001**	-
Okinawa	0.000**	0.000**	-

ADF検定とCointegration検定の結果

Granger causality検定

モデル１ $$y(t) = \sum_{i=1}^{P} a_i y(t-i) + e_0(t)$$

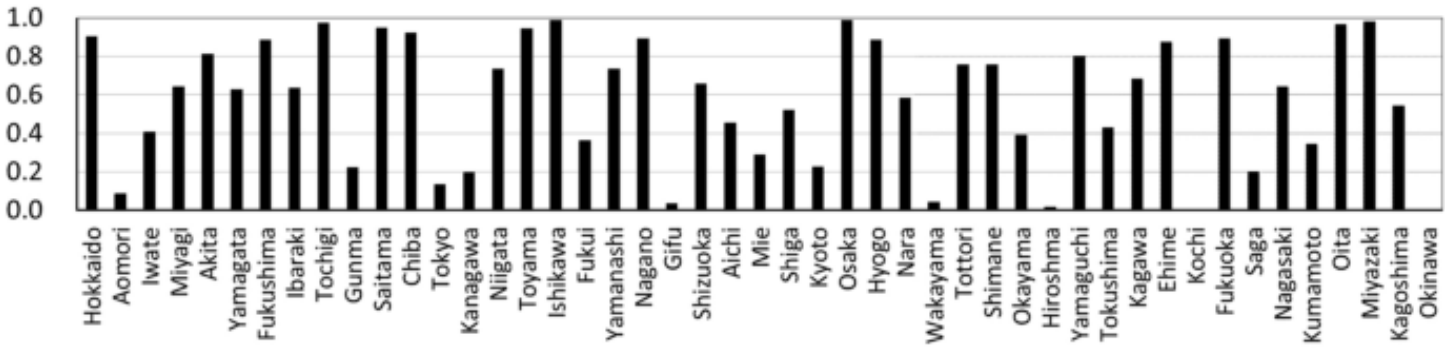
モデル２ $$y(t) = \sum_{i=1}^{P} a_i y(t-i) + \sum_{i=1}^{P} b_i x(t-i) + e_1(t)$$

モデル2のほうがモデル1より誤差が小さいことを示せば、グレンジャー因果があるといえる。

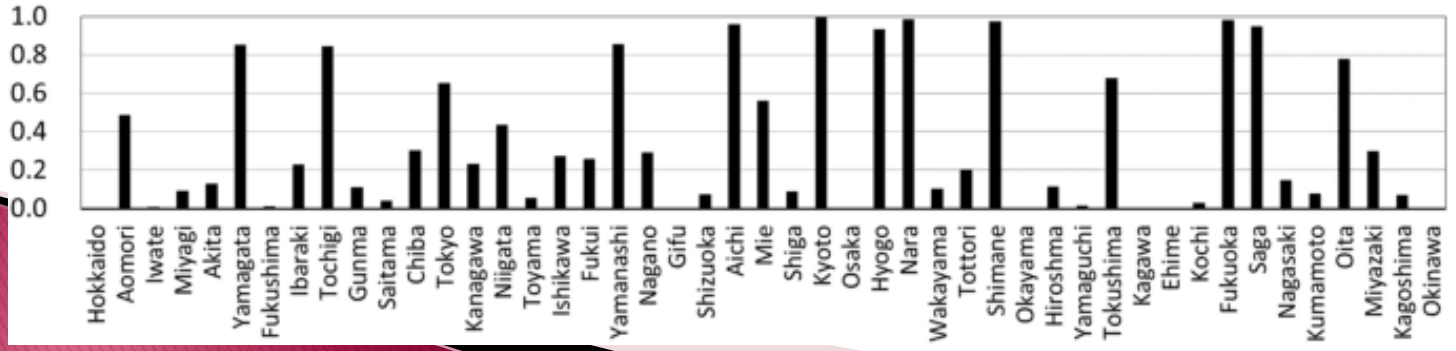
帰無仮説
・2つモデルの誤差に差がない

F検定を用いて帰無仮説を棄却する

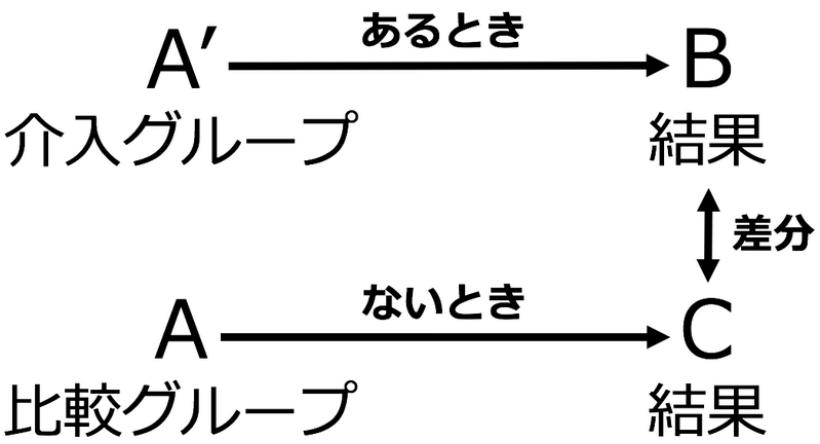
(a) p-value associated with the causality from SHR to ERN



(b) p-value associated with the causality from ERN to SHR



A/Bテスト



- 別名、ランダム化比較実験
  RCT(Randomized Controlled Trial)
- 多くの人を集めてきて、介入をしたグループと、介入をしないグループで、結果が異なるかを検定する。つまり、結果の期待値に差があるかを検定する。

A'とAが、同質なのかは可能な限り確認。できるだけ同質化

実験ができない、データ解析ではどうするのか？

「自然実験」を導入する

事例

フィールド実験：なぜ人は投票に行くのか？

投票という市民の義務を果す満足感 or 義務感があるから？

⇝ ミシガン州の有権者のうち 18 万人に投票を促す 4 種類の手紙をランダムに配布する [Gerber et al., 2008]。

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

Figure 7: Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

ビニエット実験：中国世論が強硬な対外政策を促すのか？

中国国民に中国と日本の領土紛争に関する架空のシナリオをランダムに読ませて、中国政府の行動への支持を尋ねた [Quek and Johnston, 2018]。

- 日本に対して武力を行使すると言いつつ何もしなかった $+\alpha$

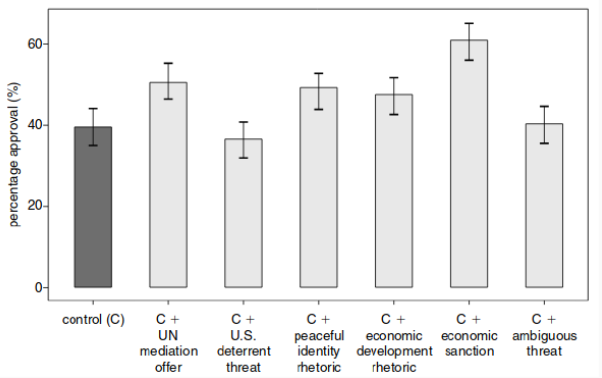


Figure 8: Public Approval for Backing Down across Experimental Groups

事例

リスト実験：人種差別を可視化する

アメリカの黒人差別の実態を把握したいが、普通に聞いても答えてくれるはずがない。

怒りを感じる項目の数を聞く
[Kuklinski et al., 1997]。

1. 政府がガソリン税を上げる。
2. プロスポーツ選手が億万長者になる。
3. 大企業が環境汚染をしている。
4. 黒人家族が隣に引っ越してくる。

統制群は最初の 3 つだけ、処置群には全ての項目を見せる。

Region	Experimental Condition: Baseline	Experimental Condition: Black Family	Estimated Percent Angry
Non-South	2.28[a] (.07) 425[b]	2.24 (.05) 461	0
South	1.95 (.06) 139	2.37 (.08) 136	42

[a]Standard error of the estimate.
[b]Number of cases.

Figure 9: Mean Level of Anger Toward A Black Family Moving in Next Door, by Region (Whites Only)

$$\mathbb{E}\left[\sum_{n=1}^{N+1} Y_{i,n}\right] - \mathbb{E}\left[\sum_{n=1}^{N} Y_{i,n}\right] = \mathbb{E}[Y_{i,N+1}]$$

コンジョイント実験：移民差別を可視化する

アメリカの移民に対する態度を把握したいが、普通に聞いても答えてくれるはずがない。

2 つの架空の移民プロフィールのうち、好ましい方を選ばせる
[Hainmueller and Hopkins, 2015]。

	Immigrant 1	Immigrant 2
Prior Trips to the U.S.	Entered the U.S. once before on a tourist visa	Entered the U.S. once before on a tourist visa
Reason for Application	Reunite with family members already in U.S.	Reunite with family members already in U.S.
Country of Origin	Mexico	Iraq
Language Skills	During admission interview, this applicant spoke fluent English	During admission interview, this applicant spoke fluent English
Profession	Child care provider	Teacher
Job Experience	One to two years of job training and experience	Three to five years of job training and experience
Employment Plans	Does not have a contract with a U.S. employer but has done job interviews	Will look for work after arriving in the U.S.
Education Level	Equivalent to completing two years of college in the U.S.	Equivalent to completing a college degree in the U.S.
Gender	Female	Male

Figure 10: Experimental Design

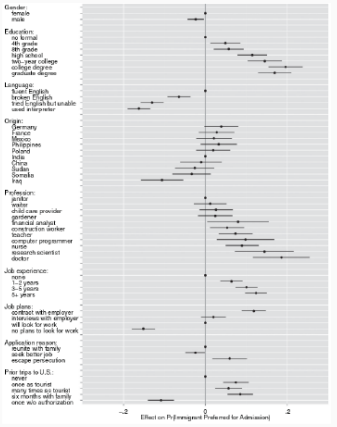


Figure 11: Effects of Immigrant Attributes on Probability of Being Preferred for Admission

自然実験

- 問題が発生している場所と、発生していない場所で違いを統計的に比較する
- 1854年8月31日、ロンドンでは、コレラの大発生がソーホーを襲った。発生の終わりまでに616人が死亡した。医師のジョン・スノーは、ポンプの周りの症例 のクラスターを明らかにした死と病気の地図を使用して、発生の原因を最も近い公共の水ポンプとして特定しました。
- 2002年、モンタナ州ヘレナでは、禁煙が実施されている間、心臓発作の発生率が40%低下したことを観察しました。法の反対者は、法の執行を6か月後に停止させることに勝ち、その後、心臓発作の割合は回復しました。

事例

自然実験：ワールドカップは世界を平和にするのか？

ワールドカップ本戦出場をナショナリズムを喚起する自然実験とみなす [Bertoli, 2017]。

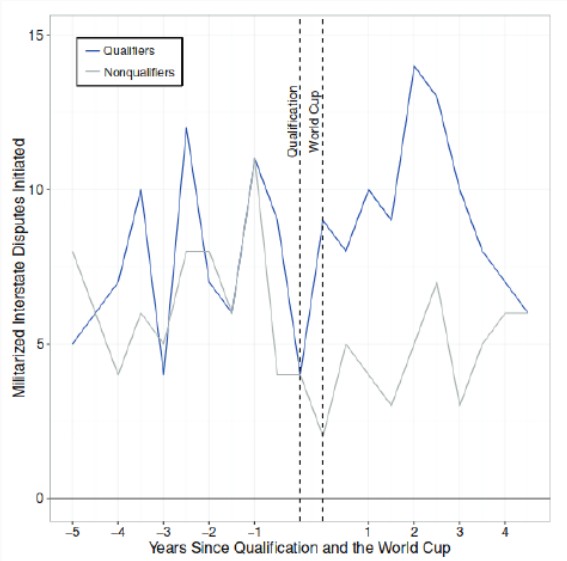


Figure 12: Comparing aggression before and after the World Cup

同質化や必要な処置のみ取り出す手法

マッチング・重み付け

仮定 2 (条件付き無作為割当)

ある属性（共変量）で条件付けると、処置は無作為に割り当てられている。

1. マッチング：処置群と統制群の間で属性が似ているものを取り出して均質化する。
2. 重み付け：重み付けをして処置群と統制群を均質化する。

例：Inverse Propensity Score Weighting

$$\hat{\tau}_{\mathrm{IPW}} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{T_i Y_i}{\hat{p}(T_i \mid X_i)} - \frac{(1-T_i)Y_i}{1-\hat{p}(T_i \mid X_i)}\right\}$$

・$\hat{p}(T_i \mid X_i)$ は傾向スコア（ある属性 X_i のもとで処置を受ける確率）

操作変数法：経済成長は内戦を抑えるのか？

直感的には正しそうだが、内戦が起こらないから経済成長が起こる可能性は十分にある。

⇝ 降雨量の増減を操作変数として、経済成長がサブサハラで内戦の発生に与える影響を推定する [Miguel et al., 2004]。

Explanatory Variable	Ordinary Least Squares (1)	(2)	(3)	(4)	(5)
Growth in rainfall, t	.055*** (.016)	.053*** (.017)	.049*** (.017)	.049*** (.018)	.053*** (.018)
Growth in rainfall, $t-1$	.034** (.015)	.032** (.014)	.028** (.014)	.028* (.014)	.037** (.015)

Figure 13: Rainfall and Economic Growth (First-Stage) Dependent Variable: Economic Growth Rate, t

Explanatory Variable	Dependent Variable: Civil Conflict ≥25 Deaths (OLS) (1)	Civil Conflict ≥1,000 Deaths (OLS) (2)
Growth in rainfall, t	−.024 (.043)	−.062** (.030)
Growth in rainfall, $t-1$	−.122** (.052)	−.069** (.032)

Figure 14: Rainfall and Civil Conflict (Reduced-Form)

操作変数法

自然実験それ自体の効果は興味ないが、うまく使うことはできないか？

仮定 3 (操作変数)

変数 Z_i が *(1)* 除外制約,$\mathrm{Cov}(Z_i, \varepsilon_i) = 0$, と *(2)* 関連性, $\mathrm{Cov}(Z_i, T_i)$, を持つとき、操作変数 *(instrumental variable: IV)* と呼ぶ。

二段階最小二乗法 (two-stage least squares: 2SLS) によって、処置の無作為な部分を切り出して ATE を求める。

1. $T_i = \alpha + \beta Z_i + \eta_i$ を推定し、予測値 $\hat{T}_i$ を求める。
2. $Y_i = \gamma + \tau \hat{T}_i + \varepsilon_i$ を予測して、$\hat{\tau}_{\mathrm{2SLS}}$ を求める。

・例として、くじ引き、天候や災害、制度上の取り決めなどがある。

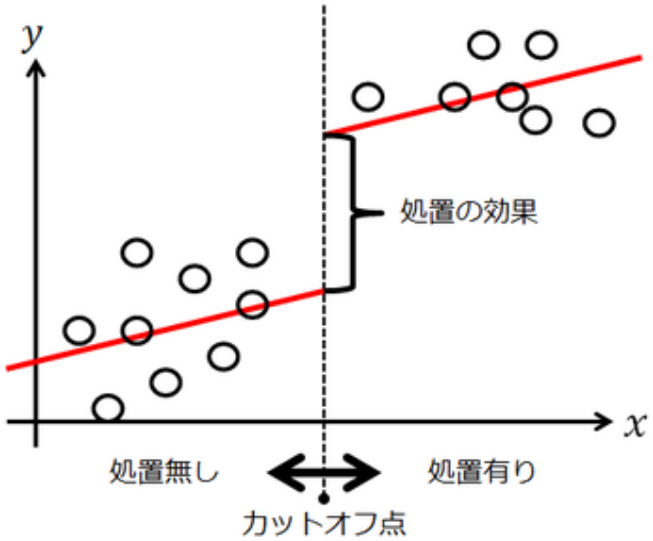
RDデザイン（regression discontinuity design）

- ビッグデータ＋自然実験＋ランダム化比較実験
- 大量のデータから似た対象を大量に見つけてきて、処置ありと、処置なしとで、処置後の結果が統計的に違うことを示す。

「奨学金の支給が、その後の成績を上昇させるか？」

- 奨学金を貰った人と、貰えなかった人で単純に成績を比較しても意味がない

成績優秀者向け奨学金が成績を全く向上させないとしても、奨学金を得た人は奨学金を貰わなかった人よりよいパフォーマンスを見せるだろう。なぜならば単純に事前に成績の良かった生徒に対して奨学金が与えられるからである。



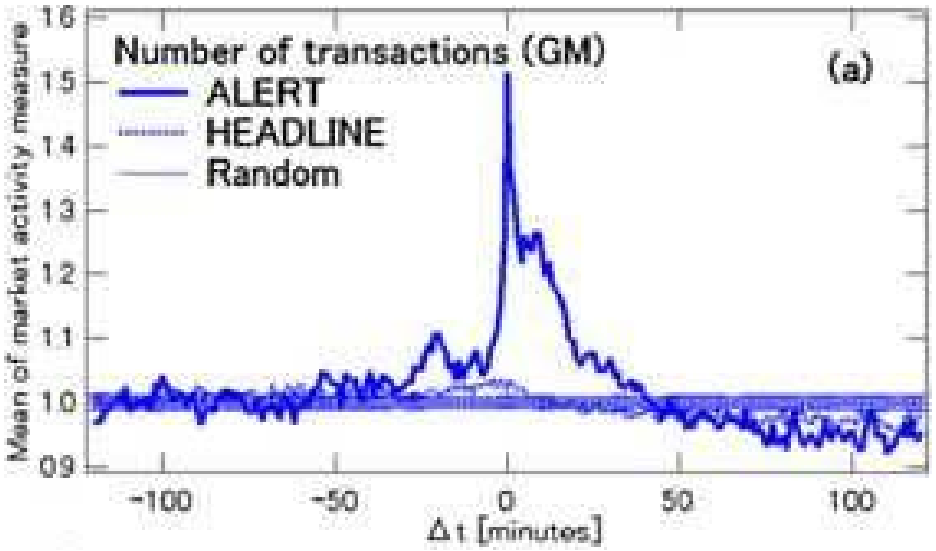
- 81点で奨学金を貰った人と、79点で貰えなかった人で成績を比較する

事前の評点が’79点の生徒と81点の生徒は非常に似ている。
もし、80点以上の全員に奨学金が与えられるのであれば、81点で奨学金を貰った人と、79点で貰えなかった人で成績を比較することで、局所的な処置効果が取り出せる。

統計的因果推論

ニュースの前後で、市場をのアクティビティを比較

新規で話題なニュースは、金融市場に影響を与える

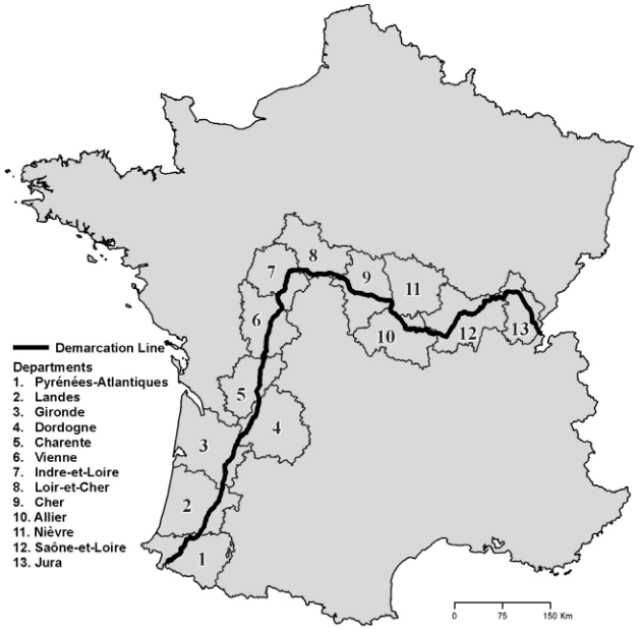


ギリギリ落選した議員と、ギリギリ当選した議員で、その後の生涯収入を比較

政治能力とは関係なく、当選が、生涯収入を上昇させる

傀儡政権が樹立された国境付近の内側と外側を比較

傀儡政権は反発を抑える。



おわりに

- 社会で起きている現象を科学的に扱おうとするとき
  - (計算社会科学以外の)人文社会科学においても多くの先行研究が存在する場合がほとんどである.
  - それらの洞察は行動ビッグデータに対する統計モデリングに役立つだけでなく, そのモデルの評価から得られた知見が, さらに計算社会科学・人文社会科学を発展させる.
- 統計的因果推論
  - 分析対象に「介入」を行った際の介入の効果を評価する枠組みである.
  - 介入の効果を最も明快に調べる方法は「介入した場合」と「介入しなかった場合」, かつ, 両者のそのほかの条件がまったく同じときに目的変数の分布を比較することである(ランダム化比較実験).

課題解決力を磨くレポート

問題の発見⇒目標の設定⇒原因の究明⇒手段の選択⇒実行と結果

- World Economic Forum
  - https://jp.weforum.org/events/world-economic-forum-annual-meeting-2019
- TED
  - https://www.ted.com/talks?language=ja

<u>レポート課題</u>

これらのWebサイトを参考に，世界が抱える社会問題を１つ選び，その問題と情報科学による解決策（計算社会科学の方法論）について，自分自身で考え，レポート用紙２,3枚（1500字から2000字まで）にまとめなさい.

レポート締め切り2/14締め切り

mizuno@nii.ac.jp