令和4年度　木更津工業高等専門学校
情報工学科　卒業研究論文

統計的因果推論による労働状態の推定

学籍番号　18-430
氏　名　花澤楓
指導教員　大枝真一

目 次

第1章　はじめに

1.1　研究背景・目的

2019 年 12 月 8 日に中国の湖北省武漢で初めて COVID-19 の症例の発生が確認された．世界では 2023 年 1 月までに 6.68 億人以上が感染し，673 万人以上が死亡している [1]. 日本政府は COVID-19 の流行を抑制するため，全国の小学校，中学校，高校の休校，4 度の緊急事態宣言や 2 度のまん延防止等重点措置，補助金の交付など多岐にわたって政策を実施してきた．また，日本だけではなく，世界中で感染症を抑制するために病理学的な処置はもちろん，各国による法令や経済政策によって様々な対策が取られてきた．また，COVID-19 の発生によって世界 GDP の成長率も 2020 年には-3.1%となり実体経済にも大きな影響を与えていることがわかる．

それらを踏まえ，Taiyo Hukai et al. (2021) [3] では日本の労働市場へ COVID-19 が与えた影響を推定し，Stefania Albanes et al. (2021) [5] ではアメリカにおける労働市場への影響を回帰分析を用いて分析している．

そこで，本研究では，労働力調査のデータを用いて，人々の労働状態，つまり雇用されているのかどうか，COVID-19 の影響はどの程度の人々を失業に追い込んだのかを統計的因果推論を用いて影響を推定する．

1.2　先行研究

1.2.1　概要

先行研究 (Taiyo Hukai et al. (2021)) [3] では，日本における LFS からの個票データ（2013 年-2022 年）を利用して，式 (1.1) の平均処置効果を Casual forest algorithm(Breiman (2001)) [4] にて求め，COVID-19 による影響を推定している．

$$\tau(x) = E[Y_{1it} - Y_{0it}|X_i = x] \tag{1.1}$$

ここで，X_i はベクトルであり回答者の年齢，性別，学歴，居住地，家族構成，雇用形式，仕事での役割，業界などの回答者の特徴を示すものである．これらを式 (2.3) と組み合わせ，条件付き期待値として求める．

また，Y_{1it} は直接 2020 年のデータから観測することができるが，Y_{0it} は観測されない．通常は，2019 年までのデータからトレンドを予測して，2020 年のデータを予測し，補完する．先行研究では，トレンドがないことを仮定して，Y_{0it} を単純に 1 年前の同じ月として分析を行っている．

1.2.2　モデル

式 (1.1) の推定には通常，線形回帰モデルが使用されるが，サンプルサイズが大きいため $\tau(x)$ が正確にはかることができない．そこで，Taiyo Hukai et al. (2021) では，ランダムフォレストの拡張版である causal forest algorithm(Breiman 2001 [4]) を $\tau(x)$ の推定に使用している．また，業界ごとや年齢ごとといったサブグループに対しても適用している．

まず，X_i から分けられたグループの平均処置効果を $G_1, \cdots, G_{20}$ で定義し，各グループにおける平均処置効果を，$\bar{\tau_l} = E[\tau(x)|x \in G_l] \quad \text{for} \quad l = 1, \cdots, 20$ とする．以下の流れの通り，これらを求める．

1. $\hat{\tau}(x))$ を推定.
2. $\hat{\tau}(x)$ に基づいてサブグループ $G_1, \cdots, G_{20}$ を作成.
3. グループごとの平均処置効果を，$\bar{\tau_l} = E[\tau(x)|x \in G_l] \quad \text{for} \quad l = 1, \cdots, 20$ より推定する．

また，$\hat{\tau}(x)$ は式 (1.2) より求める．

$$\hat{\tau(x)} = \frac{\sum_i \alpha_i(x)[Y_i - \hat{f}_Y(X_i)][I_i(2020) - \hat{f}_I(X_i)]}{\sum \alpha_i(x)[I_i(2020) - \hat{f}_I(X_i)]^2} \tag{1.2}$$

ここで，$I_i(2020)$ は 2020 年に調査されたかどうかを表す 2 値変数であり，それぞれ，$Y_i = \tau(X_i) \times I_i(2020) + f(X_i) + u_i$，$f(x) = E(Y_i|X_i = x, I_i(2020) = 0)$ を表している．

1.2.3 結果

先行研究 [3] において，LFS の回答者の従事している業界の観点から分析している． LFS の回答者のなかで，働いていないと回答した人には介護休業や育児休業など，給与を伴う休暇を取っている人や従事している環境の一時的な休業による給与の伴わない休業も含まれる．先行研究 [3] では，一時的な休業者を含む指標を “ Loose employment measure"とし，一時的な休業者を除く指標を “ Strict employment measure"としている．

ホテル，レストラン業界が Loose employment measure と Strict employment measure のそれぞれが 0.3 近くであり，COVID-19 によって 30%近くの失業者の増加影響を受けていることがわかった．これは他業界との差が顕著であり，情報系や金融業界，公務員などはほとんど影響を受けていないこともわかる．

また，サービス業に従事している従業員が顕著に影響を受けていることもわかった．以下に先行研究で判明している結果をまとめて示す．

- 若い男女，また女性が男性よりも大きく COVID-19 の影響を受けた．
- part-time job のサービス業やホテル業，レストラン業の従業員が影響を大きく受けた．
- 居住地や学歴，会社の規模は影響を与えなかった．

1.3 使用するデータ

1.3.1 概要

本研究では，The Labor Force Survey(LFS) からのカナダの家計を対象とした 2022 年 1 月から 2022 年 11 月のアンケートデータを使用する (サンプルサイズ=748390) [2].

カナダの労働力調査（LFS）は，カナダ統計局が毎月実施する調査であり，国内の労働市場の活動を推定するものである．この調査は，国，州レベルでの雇用，失業，および労働力参加に関するデータを収集している．カナダの労働市場に関する最も包括的な情報源と考えられており，政府機関，企業，研究者が意思決定や労働市場の動向を分析するために広く利用されている．この調査は世帯のサンプルに基づいており，その結果は全人口を統計的に代表するものである．

月次のデータでカナダ中の 100000 世帯から得られている．対象は，15 歳以上である．この調査は雇用情報（雇用されているかどうか）や仕事での役割，会社の規模，従事している産業，職種などの労働に関する情報に加え，回答者の年齢，性別も含まれている．

また，2023 年 1 月時点でのカナダの COVID-19 感染者数（454 万人）をカナダの総人口（3825 万人）で割ることで，カナダの人口におけるコロナ感染者数の比率（0.1186）を算出した．本研究では，LFS のデータに対して，コロナ感染者数の比率の確率でランダムにコロナ感染者かどうかを表す 2 値変数を割り当てた．これによって，擬似的にコロナ感染による労働状態への影響を推定することができる．

以下が使用するデータの代表的な変数の名前，意味，値である．データサイズは (748390, 62) である．

表 1.1: 代表的な変数名と値

Surey year	2022
Survey month	1-12
Labour force status	0：Unemployed 1：Employed
Five-year age group of respondent	1：15 to 19 years 2：20 to 24 years ... 12：70 and over
Sex of respondent	0：Female 1：Male
Martial status of respondent	1：Married 2：Living in common-law 3：Widowed 4：Separated 5：Divorced 6：Single, never married
Usual hours worked per week at main job	0.1-99 (hours)
Full- or part-time status at main or only job	0：Full-time 1：Part-time
firmsize	1：Less than 20 2：20 to 99 3：100 to 500 4：More than 500 (Number of employees)
Duration of unemployment	1-99 (weeks)
Age of youngest child	1：Less than 6 2：6 to 12 3：13 to 17 4：18 to 24
Usual hourly wages	1-999999 (canadian dollars)

1.3.2 記述統計量

表 1.2 に本研究で使用する代表的な変数の記述統計量を示す.

表 1.2: 代表的な変数の記述統計量

	平均値	標準誤差	最小値	最大値
Survey month	5.940	3.124	1.00	12.00
Labour force status	0.500	0.500	0.00	1.00
Five-year age group of respondent	5.875	3.091	1.00	12.00
Sex of respondent	0.528	0.499	0.00	1.00
Martial status of respondent	3.276	2.291	1.00	6.00
Usual hours worked per week at main job	183.7	203.6	0.00	990.0
firmsize	1.277	1.662	0.00	4.00
Duration of unemployment	7.982	17.60	0.00	99.0
Age of youngest child	0.657	1.163	0.00	4.00
Usual hourly wages	1323.8	1827.0	0.00	11298.0

第2章　統計的因果推論の枠組み

因果関係 (causality) とは，原因 (cause) と結果 (outcome) の関係であり，統計的因果推論とはデータに基づいて因果関係を明らかにすることである．つまり，原因 A が結果 B にもたらす効果を扱う．原因のことを処置 (treatment) という．

本研究では，LFS を使用して，雇用情報を結果，COVID-19 を処置として解釈する．具体的には，各家計 i の雇用情報を Y_i とし，COVID-19 に感染したことを処置 T_i とし，$T_i = 0$ で COVID-19 に感染していないとき．$T_i = 1$ で COVID-19 に感染したことを表す．処置を受ける集団を処置群，受けない集団を統制群という．これを式 (2.1) に示す．

$$Y_{it} = (1 - T_i)Y_{0it} + T_iY_{1it} = \begin{cases} Y_{0it} & \text{if} \quad T_i = 0 \\ Y_{1it} & \text{if} \quad T_i = 1 \end{cases} \tag{2.1}$$

ここで，i は各家計を表し．t は調査した月を表す．
次に示すような処置効果を求める．

2.1　処置効果 1：個体因果関係

反事実的な条件に基づく潜在的結果（potential outcome: PO）を導入して，各家計に対して，COVID-19 に感染した時と，感染していない時の結果を比べれば，COVID-19 による労働状態への影響が明らかになる．

これは，個体因果関係（ITE:individual treatment effect）と呼ばれ，式 (2.2) に示す．

$$\tau_i = Y_{1it} - Y_{0it} \tag{2.2}$$

しかし，各家計が感染しているかしていないかは，必ずどちらかの情報しか得ることができないため，τ_i は観測されず，ITE は定義できても観測も推定もできない．これは “ 因果推論の根本問題"として知られている．

2.2　処置効果 2,3：平均処置効果と処置郡の平均処置効果

平均処置効果（ATE:average treatment effect）とは，処置群と統制群の両方を含む全ての家計に対する平均効果である．処置郡の平均処置効果（ATT:average treatment effect on the treated）とは，処置を受けた集団に対する平均効果である．それぞれ式 (2.3)，式 (2.4) に示す．

$$\tau_{ATE} = E[Y_{1it} - Y_{0it}] = E[Y_{1it}] - E[Y_{0it}] \tag{2.3}$$

$$\tau_{ATT} = E[Y_{1it} - Y_{0it}|T_i = 1] \tag{2.4}$$

2.3　因果効果の識別仮定

データから式 (2.3)，式 (2.4) を識別するためには以下の仮定を満たさなければならない [8].

1. 交換性（exchangeability）
2. 一致性（consistency）
3. 正値性（positivity）
4. 相互作用なし（no interference: NI）

交換性とは，因果推論において最も重要な仮定である．全ての T=t，i について，$Y_{it} \perp\!\!\!\perp T$ が成立することである．つまり，処置 T は Y_{it} と独立であるため，t=1 と t=0 の群間では T の値を交換（exchange）しても結果が同じになることを意味する．交換性を仮定すれば，2 群の差は処置状態だけであり，実際に観察される Y の差は原因 T によってもたらされたと推論できることになる．

一致性とは，推定量が真のパラメータに確率収束するという性質ではなく，PO と観測値の関係に関する仮定である．すなわち，T=t のすべての個体について $Y_t = Y$ が成立するという仮定を意味する．

正値性とは，いずれの処置に割り当てられる確率が 0 でないことを意味する．いずれかの処置に割り当てられる確率が 0 の群があれば，潜在的結果を定義できないため正値性を仮定する必要がある．

相互作用なしの仮定は，PO が他の個体の処置状態に依存しないことであり，個体間の相互作用がないことを仮定している．

第3章　実験手法・結果

3.1　COVID-19 感染に関する仮定

本研究では，COVID-19 に関して以下の二つの仮定を考え，それぞれについて実験した．

1. COVID-19 はランダムに感染する
2. COVID-19 は各個人によって感染確率が異なる

3.2　ランダム化比較実験

ランダム化比較実験（Randomized Controlled Trial:RCT）とは，処置の割り付けが無作為であるときに，式 (2.3) を推定するものであり，これによって正確な因果効果を推定することができる．カナダの総人口における COVID-19 感染者数の比率 (454 万人/3825 万人=0.1186) の確率でランダムに割り振った．これにより，処置の割り付けの無作為化を実現した．以下が結果である．

$$\begin{aligned} \tau_{ATE} &= E[Y_{1it} - Y_{0it}] \\ &= 0.499 - 0.507 \\ &= -0.008 \end{aligned} \tag{3.1}$$

ATE は -0.008 と非常に小さく，p 値 $= 0.149 > 0.05$ より母集団の平均値に統計的有意差はない．よって，仮定 1 においては，COVID-19 による労働状態への影響はないと推定した．

3.3　傾向スコアマッチング

仮定 2 では，各個人の変数から COVID-19 に感染する確率が高い人，低い人を判断し，その確率の元ランダムに COVID-19 感染を割り当てた．具体的には先行研究 [3] を参考に，以下のような特徴をもとに判断した．

- 感染しやすい人
  - 配偶者・子供がいる
  - 週に働く時間が多い
  - 飲食・サービス業やホテル業に従事している
- 感染しにくい人
  - 失業の期間が長い

このとき，処置の割り付けが無作為ではなくなるので，RCT では正確な因果効果が推定できない．このような場合には**傾向スコアマッチング (propensity score matching:PSM)** を使用することで正確な因果効果を推定することができる．以下の手順で因果効果を推定する．

1. **傾向スコア (propensity score):**
   処置が割り付けされる確率の計算

$$\Pr(T_i = 1|X_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \tag{3.2}$$

2. **マッチング (matching):**
   $\Pr(T_i = 1|X_i)$ が近い個体で結果を比較

$$\begin{aligned}\tau &= E[Y_{1it}|T_i = 1] - E[Y_{0it}|T_i = 0] \\ &= 0.590 - 0.859 \\ &= -0.269\end{aligned} \tag{3.3}$$

p 値 < 0.05 より，母集団の平均値に有意差がある．そのため，COVID-19 によって 26.7%の負の影響があると推定した．

3.4 まとめ

本研究では，COVID-19 に関する仮定を 2 つ考え，それぞれについて実験を行なった．以下に，仮定，実験手法・結果を示す．

表 3.1: 仮定，実験手法・結果

仮定 (COVID-19 感染確率)	手法	結果 (労働状態)
ランダム	RCT	影響なし (-0.8%)
各個人により異なる	PSM	負の影響 (-26.7%)

以上より，カナダにおいて

- **COVID-19 が労働状態に与える影響は仮定により異なる．**
- COVID-19 の感染確率が各個人により異なる場合，COVID-19 に感染することで 26.7%だけ失業する可能性が高くなる．

第4章　おわりに

本研究では，COVID-19 がカナダの労働市場においてどのような影響を与えているのかを，人々の労働状態を通して推定した．COVID-19 感染に関する仮定を 2 つ想定し，それぞれについて実験した．その結果，COVID-19 がランダムに感染する場合は COVID-19 は影響を与えておらず，COVID-19 感染確率が人によって異なる，つまり COVID-19 に感染しやすい人と感染しにく人を利用可能な変数から分類し，COVID-19 を割り振った場合には，COVID-19 によって 26.7%の失業が促されたことがわかった．仮定の妥当性によるが，COVID-19 によって負の影響が与えられていることを推定した．

引用文献

[1] Our World in Data (アクセス日 2023 年 1 月 20 日 https://ourworldindata.org/explorers/coronavirus-data-explorer)

[2] Labour Force Survey: Public Use Microdata File
(アクセス日 2023 年 1 月 20 日 https://www150.statcan.gc.ca/n1/pub/71m0001x
/71m0001x2021001-eng.htm)

[3] Taiyo Fukai, Hidehiko Ichimura, Keisuke Kawata, (2021), Describing the impacts of COVID-19 on the labor market in Japan until June 2020, The Japanese Economic Review

[4] L.(2001). Random forests. Machine leaning

[5] Stefania Albanesi, Jiyeon Kim, (2021), Effects of the COVID-19 Recession on the US Labor Market: Occupation, Family, and Gender

[6] 高橋将宜, 2022, 統計的因果推論の理論と実装, 共立出版

[7] 安井 翔太, 2020, 効果検証入門〜正しい比較のための因果推論/計量経済学の基礎, 技術評論社

[8] 大久保 将貴 (2019), 因果推論の道具箱,J-STAGE