格差/因果/比較分析 (ver 0.2)

Balancing weights を軸とした手法整理と機械学習の活用

川田恵介

2025-02-19

Table of contents

Preface

定量的な比較分析の方法を、R での実装とともに紹介します。

比較分析は、社会における「グループ間の違い」を明らかにすることを目標とします。本ノートでは、データ上のある変数 D 間での、別の変数 Y の平均値の差を推定する方法を紹介します。例えば性別 (D) 間での賃金 (Y) の平均格差を推定します。

比較分析は、社会/市場を理解するための方法として、中核的な位置を占めています。例えば、ある職業訓練プログラム (D) が就業確率や就業後の賃金 (Y) に与える因果的効果を推定を試みます。このような研究課題では、異なる職業訓練プログラムへの参加者間で、就業状態や賃金の比較が求められます (Behaghel, Crépon, and Gurgand 2014; Kallus 2023)。

比較分析は格差研究でも重要です。格差の「世代間」移転を推定するためには、両親の経済/社会状態 (D) と子供の経済/社会状況 (Y) との関係性を推定します。例えば、母親の学歴 (大学卒/高校卒など) グループ間で、その子供の教育達成度賃金を比較します。

格差の” 要因” を考察するために行われる分解分析も、比較研究の一種です[*1]。Vafa, Athey, and Blei (2024) では、同一の職業経験を有する男女間での賃金格差を推定しています。男女間での単純な賃金格差と比較することで、職業経験の性別間分断が、賃金格差に与えている影響について示唆を得ることができます。

以上のような比較研究では、しばしば「背景属性」のバランスが求められます。例えば 2022 年と 2023 年の中古マンション市場の取引価格を比較したいとします。最も単純な比較方法は、平均取引価格を 2023 年と 2022 年で単純比較することですが、物件の他の特徴 (部屋の広さ、駅からの距離など) も同時に変化していることが予想されます。本ノートでは、「もし物件の他の特徴が変化しなかったときの」2023 年と 2022 年で平均取引価格がどの程度変化額を推定する方法を紹介します。

本ノートの中心的なコンセプトは、Balancing Weight です。Balancing Weight は、以下のアプローチを統合的整理できる極めて有益な概念です。

[*1] Opacic, Wei, and Zhou (2023) が、包括的な紹介を行っています。

- 重回帰 (OLS) や Penalized Regression による調整 (Chattopadhyay and Zubizarreta 2023; Bruns-Smith et al. 2023)
- 傾向スコア (Propensity Score) の活用 (Imai and Ratkovic 2014; Chernozhukov et al. 2018, 2022)
- Entropy Weight (Hainmueller 2012) や Stable Weight (Zubizarreta 2015)、Energy Weight (Huling and Mak 2024) による調整
- 機械学習などを用いた Auto Double/Debiased Machine Learning/Augmented Balancing Weights (Chernozhukov, Newey, and Singh 2022a, 2022b; Bruns-Smith et al. 2023)

簡易な入門としては Chattopadhyay and Zubizarreta (2024) 、詳細な入門としては Chattopadhyay, Hase, and Zubizarreta (2020) , Ben-Michael et al. (2021) などを参考にしてください。

本ノートの構成は以下のとおりです。

- Chapter 1：本ノートの Estimand である「バランス後の比較」を定義します。
- Chapter 2：本ノートの中核概念である Balancing Weight を紹介する準備として、荷重 (Weight) を定義します。
- Chapter 3：Balancing Weight を定義します。また Balancing Weight の直感的な算出方法が、利用できない状況が多いことを指摘します。
- Chapter 4：より幅広い状況で算出できる近似的な Balancing Weight を紹介します。また標準的な OLS 推定が、暗黙のうちに近似的な Balancing Weight を算出した、バランス後の比較と解釈できることを示します。
- Chapter 5：近似的な Balancing Weight を、明示的な最適化問題として算出する方法を紹介します。Entropy weight (Hainmueller 2012) と呼ばれる代表的な方法は、その計算効率や分析の透明性の高さから、幅広く用いられています。
- Chapter 6：OLS よりもデータ主導のアプローチを紹介します。残差回帰に機械学習を補助的に用いることで、近似的な Balancing Weight を算出した、バランス後の比較を行うことができます。OLS とは異なり、事例数が十分大きかえれば、「母集団上で、完全な Balance を達成した後の比較分析」、の優れた推定値と見做せることを紹介します。

第 1 章

バランス後の比較

典型的なバランス後の比較 (Balanced Comparison) 分析では、グループ D の間で、X についての**差を解消した後に**、Yについての平均差を推定します[*1]。このような比較は、格差分析や因果効果の肝となります。

まず実例から紹介します。

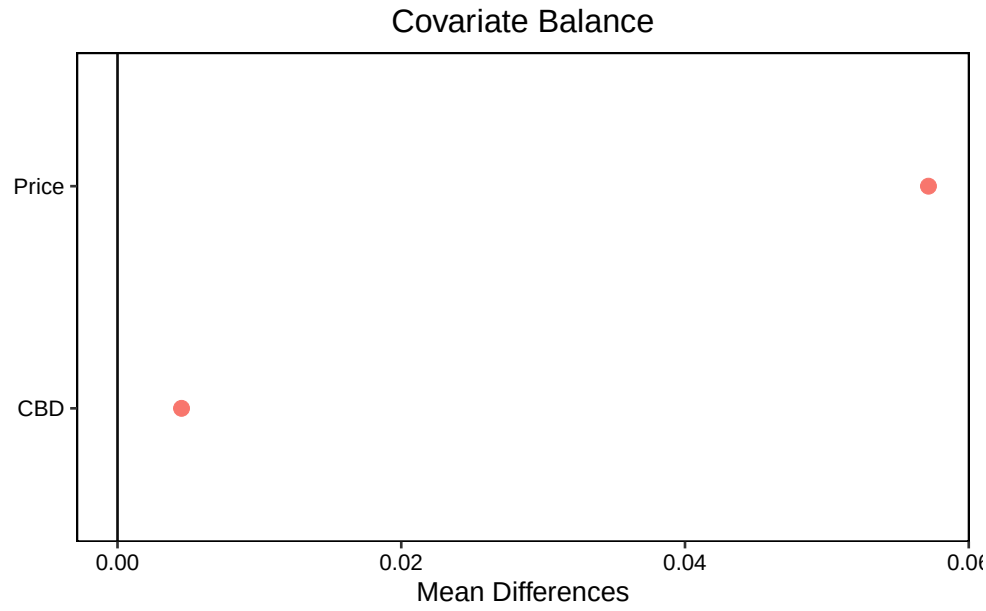
1.1 不動産市場の年次比較

東京 23 区の中古マンション市場において、2022 年と 2021 年の取引価格と立地 (中心 6 区 (CBD; 港区、中央区、文京区、千代田区、渋谷区、新宿区)、かそれ以外か) について、平均的な差を図示します。

[*1] Gelman, Hullman, and Kennedy (2024)

Example

平均価格 (100 万円)	CBD	取引年	事例割合
37.748	0	2021	0.784
60.474	1	2021	0.216
39.150	0	2022	0.779
64.814	1	2022	0.221



2022 年の平均取引価格は、2021 年に比べて上昇しており、不動産市場における価格上昇が続いているように見えます。しかしながら、同時に CBD の物件割合も増加しています。一般に CBD に立地する物件の方が、高い取引価格が予想されます。このため物件の立地の変化によって、取引価格の上昇が” 底上げ” されている可能性があります。

もし中心 6 区の物件割合が不変であった場合、平均取引価格にどのような差が残るでしょうか？ このような問いに対して、バランス後の比較分析は回答できます。

以下では実際のデータを用いて、取引年と立地 (CBD であれば 1、それ以外であれば 0) ごとに、平均取引価格を示しています。また各取引年における取引事例の立地割合も算出しています。

繰り返し期待値の法則を用いると 2022 年と 2021 年の平均取引価格の差を算出できます。

❗ 繰り返し期待値の法則

- 変数 Y, D と他の変数 X について、Yの平均値 は以下のように書き換えられる。

$$(D=d)\text{における}Y\text{の平均値}$$

$$=(D=d, X=x_1)\text{を満たす事例の}Y\text{の平均値}\times(D=d, X=x_1)\text{を満たす事例の割合}$$

$$+..$$

$$+(D=d, X=x_L)\text{を満たす事例の}Y\text{の平均値}\times(D=d, X=x_L)\text{を満たす事例の割合}$$

- ただし X がとりえる値は、$x_1, .., x_L$ とする。

Example に適用すると、2022 年の平均取引価格は以下のように計算できます。

$$2022\text{年の平均価格}$$

$$= \underbrace{64.814}_{CBD\text{における平均価格}} \times \underbrace{0.221}_{CBD\text{に立地する事例の割合}} + \underbrace{39.150}_{CBD\text{以外における平均価格}} \times \underbrace{0.779}_{CBD\text{以外に立地する事例の割合}}$$

$$= 44.810$$

2021 年の平均取引価格も同様に計算できます。

$$2021\text{年の平均価格} = 60.474 \times 0.216 + 37.748 \times 0.784 = 42.658$$

2022 年と 2021 の平均差を計算すると、2022 年の平均取引価格の方が、2.153 ほど高くなっていることが確認できます。

繰り返し期待値の法則自体は、シンプルな計算ルールですが、「X のバランス」の意義を明確にします。繰り返し期待値の法則から、平均価格の違いは、同じ立地内での平均取引価格の違いと立地の割合の違いによって生じることがわかります。22021 年と比べると、2022 年に取引された物件の中で、CBD の物件割合が 21.6 % から 22.1 % に上昇しています。一般に CBD の方が、物件の価格が高い傾向にあります。このため、CBD の物件割合の違いが、平均取引価格上昇の一因となっている可能性があります。

平均価格	CBD	取引年	事例割合	ターゲットとなる割合
37.748	0	2021	0.784	0.750
60.474	1	2021	0.216	0.250
39.150	0	2022	0.779	0.750
64.814	1	2022	0.221	0.250

バランス後の比較分析では、「取引されている物件に占める CBD の割合が、変化しなかった場合」の平均取引価格の変化を推定します。より一般的には、「X についての差を解消した状態で」D 間で Y を比較します。

1.2　バランス後の平均差

本ノートにおいて、バランス後の比較分析の一つである「バランス後の平均差」を推定します。

❗ バランス後の平均値

$$(D = d)\text{における}Y\text{の平均値}$$

$$= (D = d, X = x_1)\text{を満たす事例の}Y\text{の平均値} \times (X = x_1)\text{についてのターゲットとなる割合}$$

$$+..$$

$$+(D = d, X = x_L)\text{を満たす事例の}Y\text{の平均値} \times (X = x_L)\text{についてのターゲットとなる割合}$$

- ターゲットとなる割合は、D の値にかかわらず同じ値として、研究者が定める。

バランス後の平均値の計算例として、以下ではターゲットを取引年にかかわらず、「CBD が 0.25, その他が 0.75」と設定し、Example に適用します。

ターゲットのもとでは、取引年にかかわらず、CBD とそれ以外に立地する事例の割合は一定となります。このため平均価格の変化を生み出す要因から、事例割合の変化を排除できます。

事例割合をターゲットに差し替えて、繰り返し期待値の法則を適用すると、バランス後の平均価格は以下のように計算できます。

2022年のバランス後の平均価格

$$= 64.814 \times \underbrace{0.25}_{CBDについてのターゲット} + 39.150 \times \underbrace{0.75}_{CBD以外についてのターゲット} = 45.566$$

$$2021年のバランス後の平均価格 = 60.474 \times 0.25 + 37.748 \times 0.75 = 43.4295$$

バランス後の平均差は、2.1365 であり、バランス前 (2.153) に比べて縮小しました。

1.2.1 Overlap の仮定

バランス後の平均差が計算できる前提条件は、Overlap の仮定 (Positivity の仮定とも呼ばれます) が成り立っていることです。

> **Important 1: Overlap の仮定**
>
> - ターゲットとなる割合が 0 よりも大きい X の組み合わせについて、$D = 1$ の事例も $D = 0$ の事例も、**両方**存在する:
>
> $$1 > \Pr[D = d|X] > 0$$
>
> ただし $\Pr[D = d|X]$ は $(X = x)$ 内での $D = d$ の割合 ($D = 1$ の割合)

Overlap が成り立っていない場合、全ての D について平均値が計算できず、厳密なバランス後の比較は**根本的に**不可能です。

例えば、教育経験 $(= X)$ をバランスさせた男女間 $(= D)$ での賃金格差を推定したいとします。もしデータにおける男女間での教育経験の分断が極めて大きい場合、大学卒以上の女性は存在しない可能性があります。この場合、$X =$ 大学卒 の女性割合は 0 であり、大学卒内で男女間比較は不可能です。このため大学卒について、ターゲットとなる割合を”0”としない限り、バランス後の比較は不可能です。

1.3 他の実例

バランス後の比較分析の応用例は、大量に存在します。以下では実務においてよく利用されている指標を紹介します。

1.3.1　合計特殊出生率

バランス後の比較分析は、社会/政策分析において幅広く利用されています。代表例は、合計特殊出生率の国家間/時代間比較です。

出生数の動向を把握する上で、新生児数を年次や国家間比較は、有益だとみなされてきました。合計特殊出生率は、成人の年齢構造の違いをバランスさせるために利用されている指標です。単純な出生率（一年間に生まれた子供の数/女性の数）は、成人の年齢構造の影響を強く受ける可能性があります。社会において高齢者の比率が高まれば、出生率は低下することが予想されるからです。対して合計特殊出生率は、「仮に年齢構造が同じであった場合」の出生率を、以下の方法で推定しています

$$\frac{\text{15歳の女性が産んだ子供の数}}{\text{15歳の女性の数}} + .. + \frac{\text{49歳の女性が産んだ子供の数}}{\text{49歳の女性の数}}$$

合計特殊出生率の計算においては、15 歳から 49 歳の女性の年齢 $(= X)$ 層が同じ割合になることを目指しています。すなわちターゲットなる割合は、15 歳から 49 歳までについて 1/34、それ以外の層については”0” となります。

1.3.2　既存店ベースの比較

バランス後の比較は、企業の経営戦略を考える上でも用いられます。

小売や飲食/宿泊業などでは、しばしば既存店に絞った上での、売上比較がなされます。例えば、あるコンビニチェーンで、店舗あたりの平均売り上げが 1000 万円増大したとします。同時に去年から今年にかけて、新規出店も大きく増加したとします。新規店の方が売上が高くなる傾向がある場合、新規店割合の違いが、平均売上の上昇をもたらした可能性があります。

既存店割合をバランスさせるシンプルな方法として、既存店のみに絞った平均売上を比較がよく行われます。このような分析では、既存店比率は全ての年について 100% となり、完全なバランスが達成されます。

既存店ベースの比較におけるターゲットは、新規店については 1、新規店以外については 0 となります。

1.4　応用上の課題

バランス後の比較は、シンプルな枠組みです。上記の例の通り、大規模なデータを用いて、少ない X をバランスさせるのであれば、単純な計算でバランスできます。

しかしながら、多くの応用では複数の X を同時にバランスさせることが求められます。このような応用では、同じ属性を持つ事例数が少なくなり、実際の計算は困難になります。

多くの応用で活用できる、より実践的な推定方法が求められます。このような推定手法は、Balancing Weight (Chapter 3) を実質的に推定していると解釈できます。Chapter 2 では Balancing Weight を理解するための準備として、一般的な Weight を紹介します。

第 2 章

Weight

Balancing Weights を導入する準備として、より一般的な概念である Weights を紹介します。Weights は、データ上の事例の分布を、統計的な処理によって、変化させるために用いられます。

以下、サンプリングの偏りへの対応を例とします。前章のバランス後の比較分析と議論の多くが類似している点に注意してください。

2.1 データ上の平均値

今、ある「不動産研究所」が調査員を千代田区、文京区、板橋区に派遣し、中古マンションの取引事例を収集したとします。各調査員は、全く同じ数の事例を収集します。

以下では、立地 (District) ごとに、平均取引価格とデータ全体に対する事例数の割合をまとめています。

全ての区について、同数の事例が収集されていることに注意してください。

繰り返し期待値の法則を用いると、この情報のみからデータ全体の平均取引価格は計算できます。

- 平均取引価格 は

Example

事例の割合	平均価格	District
0.333	66.5	千代田区
0.333	47.3	文京区
0.333	29.5	板橋区

$$
\begin{aligned}
&= \underbrace{66.5}_{\text{千代田区の事例の平均取引価格}} \times \underbrace{0.333}_{\text{千代田区の事例の割合}} \\
&+ \underbrace{47.3}_{\text{文京区の事例の平均取引価格}} \times \underbrace{0.333}_{\text{文京区の事例の割合}} \\
&+ \underbrace{29.5}_{\text{板橋区の事例の平均取引価格}} \times \underbrace{0.333}_{\text{板橋区の事例の割合}} \\
&= 47.8
\end{aligned}
$$

繰り返し期待値の法則から、データ上の平均値は、X についてのサブグループ内での平均値とサブグループの割合の掛け算の総和となります。このため、もしサブグループの割合が研究関心から乖離している場合、サブグループ内での平均値が妥当な値であったとしても、ミスリードな平均値が計算されてしまいます。

2.2 ターゲット上の平均値

今研究関心は、「もし**実際の取引履歴をすべて収集したデータ**を用いて計算された平均取引価格」、であるとします。このような平均値を計算したい仮想的なデータを、バランス後の比較分析と同様に、**ターゲット**と呼びます。

今、ターゲットの取引割合は、千代田区が 0.161、文京区が 0.33、板橋区が 0.509 であることが判明しているとします。もしデータ上の各区の取引割合を、ターゲットの取引割合と一致させた場合、平均値はどのように変化するでしょうか？

Example における調整された平均取引価格は、以下のように算出されます。

$$
\begin{aligned}
&= \underbrace{66.5}_{\text{千代田区の事例の平均取引価格}} \times \underbrace{0.161}_{\text{千代田区の事例の割合}} \\
&+ \underbrace{47.3}_{\text{文京区の事例の平均取引価格}} \times \underbrace{0.33}_{\text{文京区の事例の割合}} \\
&+ \underbrace{29.5}_{\text{板橋区の事例の平均取引価格}} \times \underbrace{0.509}_{\text{板橋区の事例の割合}} \\
&= 41.3
\end{aligned}
$$

データ上の平均値は 47.8 であったので、過大であったことがわかります。これは、平均取引価格が高い傾向にある千代田区の物件割合が、現実の取引割合 (0.161) よりも、データ上の割合 (0.333) が過大であることに起因します。

2.2.1 バランス後の平均との類似性

以上の議論は、バランス後の平均値と本質的には同じものです。ターゲットとデータ上の分布が乖離しているため、平均値と調整された平均値は乖離しています。唯一の違いは、バランス後の平均値を定義する際には、ターゲットを D の値に依存しないように設定する必要があり、Overlap の仮定 Important 1 に注意する必要がある点のみです。

このため以下の加重平均を用いた調整された平均値の計算方法は、バランス後の平均値を求める際にも利用できます。

2.3 加重平均値

調整を行うための有力な方法は、加重平均値 (Weighted mean) を計算することです。一般に加重平均値は以下のように定義されます。

加重平均値

- 調整された平均値

$$= (\omega \times Y) \text{の平均値}$$

- $\omega =$ 加重 (Weight)、以下の制約を置く
  - ωの平均値 $= 1$

Weight は各事例の Y の値が、最終的な平均値に反映される度合いをコントロールします。例えば、もし $\omega = 0$ であれば、その事例は平均値の計算に一切反映されません。

Weight は、データとターゲットにおける X の分布を揃えるように設定されます。すなわち

$$\omega \times \text{データ上の} X \text{の割合} = \text{ターゲットとなる} X \text{の割合}$$

を達成するように ω を算出します。両辺を事例割合で割ると、

$$\omega = \frac{\text{ターゲットとなる割合}}{\text{データ上の割合}}$$

平均価格	District	ターゲットとなる割合	事例の割合	Weight
66.5	千代田区	0.161	0.333	0.483
47.3	文京区	0.330	0.333	0.991
29.5	板橋区	0.509	0.333	1.529

となります。すなわちターゲットよりも過大に収集されているグループは小さめに、ターゲットよりも過小なグループは大きめに反映させます。

Example に適用すると、以下となります。

加重平均値は以下のように算出できます。

$$\Big[\underbrace{\underbrace{44}_{\text{取引価格}} \times \underbrace{0.483}_{\text{荷重}} + 70.0 \times 0.483 + ..}_{\text{千代田区の事例}} \underbrace{+ 75.0 \times 0.991 + ..}_{\text{文京区の事例}}$$

$$\underbrace{+ 59.0 \times 1.529 + ..}_{\text{板橋区の事例}}\Big] \text{の平均} = \underbrace{41.3}_{\text{調整された平均値}}$$

第 3 章

Balancing Weight

D 間での X の分布をバランスを達成する実用的な手法は、数多く提案されています (Chattopadhyay, Hase, and Zubizarreta 2020; Bruns-Smith et al. 2023)。このような手法を整理し、活用していくためには、**Balancing weight** という概念を導入することが有益です。

Balancing weight は、前章で導入した Weight の一種であり、D 間での X の分布の乖離を調整するために用いられます。

3.1 Balancing Weight の定義

Balancing Weight

- Balancing weight $\omega(x, d)$ は、D 間での X の分布の乖離を調整するために導入され、以下のように定義する。

$$
\begin{aligned}
&D = 1\text{における事例割合} \times \omega(x, 1) \\
&= D = 0\text{における事例割合} \times \omega(x, 0) \\
&= \text{ターゲットとなる割合}
\end{aligned}
$$

- 定義式を変形すると

$$
\omega(x, d) = \frac{\text{ターゲットとなる割合}}{(D = d)\text{グループにおける割合}}
$$

  - ターゲットに比べて過大な事例割合が過大なグループに対しては小さい、過小なグループに対しては大きな Weight を付与する。

ターゲットは、原理的には研究者が指定する必要があります。代表的なものは、データ全

平均価格	CBD	取引年	事例割合	ターゲットとなる割合	Balancing Weight
37.748	0	2021	0.784	0.782	0.997
60.474	1	2021	0.216	0.218	1.010
39.150	0	2022	0.779	0.782	1.003
64.814	1	2022	0.221	0.218	0.989

体における X の分布です[*1]。

先のデータに適用すると、以下のような Balancing Weight が計算されます。

2022 年と 2021 年を結合したデータ全体のうち、CBD に立地する物件割合は 21.8% 、それ以外が 78.2% であったので、ターゲットして設定しています。

Balancing weights を用いると、バランス後の平均値は以下のように計算できます

$$2021\text{年のバランス後の平均値} = 2021\text{年の}(\omega(X_i, 2021) \times Y_i)\text{の平均値}$$

$$2022\text{年のバランス後の平均値} = 2022\text{年の}(\omega(X_i, 2022) \times Y_i)\text{の平均値}$$

3.2 母集団の推定方法

X の組み合わせの種類に比べて、十分な事例数が存在するのであれば、Balancing weight は、データ上での X の割合を用いて計算できます。

この方法は Exact Matching や Stratified Estimation (Wager 2024) として知られる方法による推定結果と完全に一致します。例えば Exact Matching は、MatchIt package (Stuart et al. 2011) などを利用して実装できます。

さらに「母集団上で、Exact Matching や Stratified Estimation を行った場合の結果」について、推論することもできます。事例全体の数が十分にある場合、信頼区間を近似計算できるためです[*2]。

3.3 R による実践例

- 以下のパッケージを使用
  - readr (tidyverse に同梱): データの読み込み

[*1] 因果推論の文脈では、平均効果 (Average Treatment Effect) と呼ばれています。
[*2] 詳細に関心がある読者は、Wager (2024) の 2 章等を参照ください。

- matchit: Exact matching を含む多様な Matching を実装
  - mannual
- estimatr: Robust stnadard error を計算する
- gt: 見やすいテーブルを出力

3.3.1 準備

データを取得します。D として、中心 6 区かそれ以外で、1/0 となる変数を定義します。シンプルな比較分析について信頼区間は、データ分割は不要です。

```
Data = readr::read_csv("Public.csv") # データ読み込み

Data = dplyr::mutate(
  Data,
  D = dplyr::if_else(
    LargeDistrict == "中心 6 区",1,0
  )
) # D の定義
```

3.3.2 Balancing Weight

MatchIt パッケージ内の matchti 関数を用いて、Balanced weights を計算します。例えば部屋の広さ (Size) についてバランスした平均取引価格とその信頼区間は、以下で計算できます。

```
Match = MatchIt::matchit(
  D ~ Size, # D ~ X を指定
  Data, # 用いるデータの指定
  method = "exact", # Balanced weight を計算するために、exact matching を実行
  target = "ATE" # サンプル全体の X の分布をターゲット
  )

DataWeight = MatchIt::match.data(
  Match,
  drop.unmatched = FALSE # Balance weight が計算できない事例も含む
  ) # Balance weight を含んだデータを生成

Match # Balance weight の特徴を表示
```

```
A `matchit` object
 - method: Exact matching
 - number of obs.: 6378 (original), 6378 (matched)
 - target estimand: ATT
 - covariates: Size
```

number of obs. において、元々の事例数 (6378) と balanced weight を計算できた事例数 (6378) を表示しています。事例数は減少しておらず、Balancing Weight を計算できない事例が存在しないことを確認できます。

バランス後の平均差を推定すると以下となります。

```
estimatr::lm_robust(
  Price ~ D,
   DataWeight,
   weights = weights # Balancing weights を使用
   )
```

	Estimate	Std. Error	t value	Pr(>|t|)	CI Lower	CI Upper	DF
(Intercept)	36.83243	0.3625188	101.6014	0.000000e+00	36.12177	37.54309	6376
D	22.14787	1.2648841	17.5098	4.392756e-67	19.66827	24.62746	6376

単純比較の結果は以下であり、大きく異なることが確認できます。

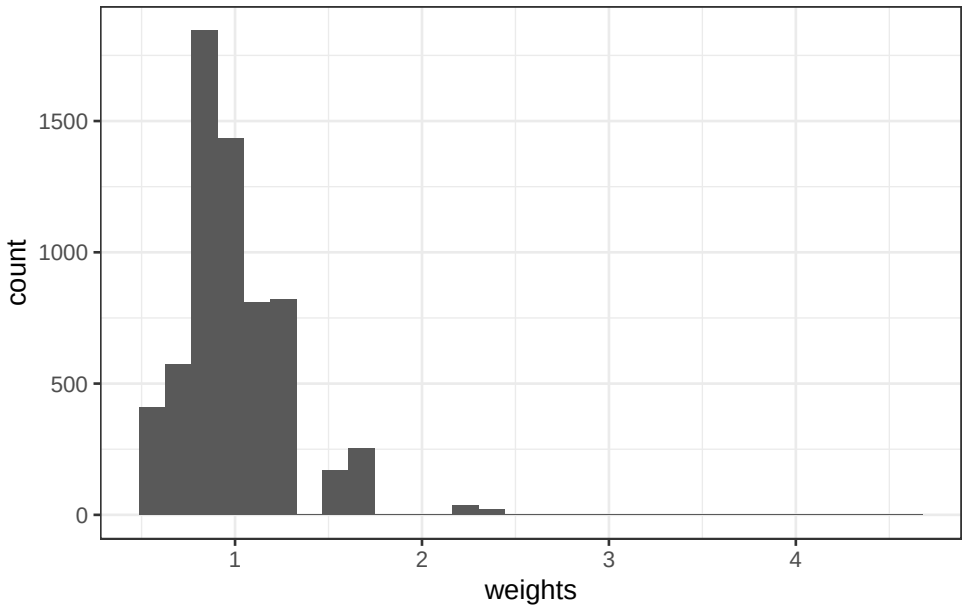
```
estimatr::lm_robust(
  Price ~ D,
  DataWeight)
```

	Estimate	Std. Error	t value	Pr(>|t|)	CI Lower	CI Upper	DF
(Intercept)	38.03972	0.3182179	119.5399	0.000000e+00	37.41591	38.66354	6376
D	20.94057	1.2529064	16.7136	2.084367e-61	18.48446	23.39669	6376

Balancing Weight の分布は、以下のように確認できます。

```
DataWeight |>
  ggplot(
    aes(
      x = weights
    )
  ) +
  geom_histogram() +
  theme_bw()
```

Price	District	Size	Tenure	TradeYear	StationDistance	LargeDistrict	D	weights	subclass
86	大田区	120	32	2021	6	その他	0	4.651185	20
110	世田谷区	120	32	2021	3	その他	0	4.651185	20
78	世田谷区	120	26	2022	4	その他	0	4.651185	20
72	文京区	115	33	2021	7	その他	0	2.325592	19
230	台東区	105	2	2021	4	その他	0	2.325592	17
88	江東区	110	17	2022	6	その他	0	2.325592	22
54	目黒区	110	53	2022	3	その他	0	2.325592	22
50	目黒区	105	20	2022	13	その他	0	2.325592	17
49	大田区	105	19	2021	21	その他	0	2.325592	17
70	大田区	105	30	2021	7	その他	0	2.325592	17



一部、非常に大きな値をとるグループがあります。Weights が大きい上位 10 事例は以下のとおりです。

```
DataWeight |>
  arrange(-weights) |>
  head(10) |>
  gt::gt()
```

3.4　応用上の課題

Exact matching や Stratified Estimation は、非常に直感的な推定方法ですが、X の組み合わせが増えると、実行不可能です。例えば X に、両親の年収や資産などの連続変数が含まれている場合は、X の組み合わせが非常に大きくなり、Balancing weights を計算することは事実上不可能となります。

例えば、Size, Tenure (築年数), StationDistance (駅からの距離)、District (立地) について、Balancing Weights は、以下で計算を試みることができますが、エラーが表示されます。

```
Match = MatchIt::matchit(
  D ~ Size + Tenure + StationDistance, # D ~ X を指定
  Data, # 用いるデータの指定
  method = "exact", # Balanced weight を計算するために、exact matching を実行
  target = "ATE" # サンプル全体の X の分布をターゲット
  )
```

このエラーは、これら 4 つの属性が全く同じ事例は、中心 6 区かそれ以外のどちらからしか存在しないことを意味しています。

もし District のバランスを諦めれば、Balancing Weight 自体は計算できます。

```
Match = MatchIt::matchit(
  D ~ Size + Tenure + StationDistance, # D ~ X を指定
  Data, # 用いるデータの指定
  method = "exact", # Balanced weight を計算するために、exact matching を実行
  target = "ATE" # サンプル全体の X の分布をターゲット
  )

Match
```

```
A `matchit` object
 - method: Exact matching
 - number of obs.: 6378 (original), 1702 (matched)
 - target estimand: ATT
 - covariates: Size, Tenure, StationDistance
```

ただし Weight を計算できたのたのは、6378 事例の内、1702 事例のみです。このような大幅な事例減少は、分析のターゲットとなる分布からデータを大きく乖離させ、ミスリードな推定結果を生み出す要因となります。

この問題を解決するために、次節以降で紹介する、OLS (Chapter 4) や機械学習 (Chapter 6) などを用いた「近似的なバランス法」の活用が有用です。

第 4 章

OLS による特徴のバランス

X の組み合わせが多く、Balancing weights を計算することが困難な場合、X の分布を近似的なバランスが有力です。

本節では代表的な統計手法である OLS（重回帰）が、近似的な Balancing weights を暗黙のうちに計算する手法であることを紹介します。OLS は分布の特徴を、研究者が定める定式化に応じて、柔軟に Balance できることが利点ですが、予期せぬ挙動を示しやすいことに注意が必要です。

4.1 OLS による平均値のバランス

近年の研究により、線型モデルの OLS 推定は、近似的な Balance を達成することが確認されています (Chattopadhyay and Zubizarreta 2023)。本ノートでは、$D = \{0, 1\}$ を前提とし、その議論の骨子を紹介します。

❗ OLS の性質 [@chattopadhyay2023implied]

- $Y \sim D + X_1 + .. + X_L$ を OLS で推定し算出される D の係数値は、以下の方法で計算される値と完全に一致する

1. すべての X_l について、以下を満たす $\omega(x, d)$ を探す。

$$D = 1\text{における}(\omega(x, 1) \times X_l)\text{の平均値}$$

$$= D = 0\text{における}(\omega(x, 0) \times X_l)\text{の平均値}$$

2. 1. を満たす $\omega(x, d)$ の中で、最小の分散を持つ $\omega(x, d)$ を Balancing Weights とする

3. β_D を以下のように計算する。

$$\beta_D = D = 1\text{における}(\omega(x,1) \times Y)\text{の平均値}$$

$$-D = 0\text{における}(\omega(x,0) \times Y)\text{の平均値}$$

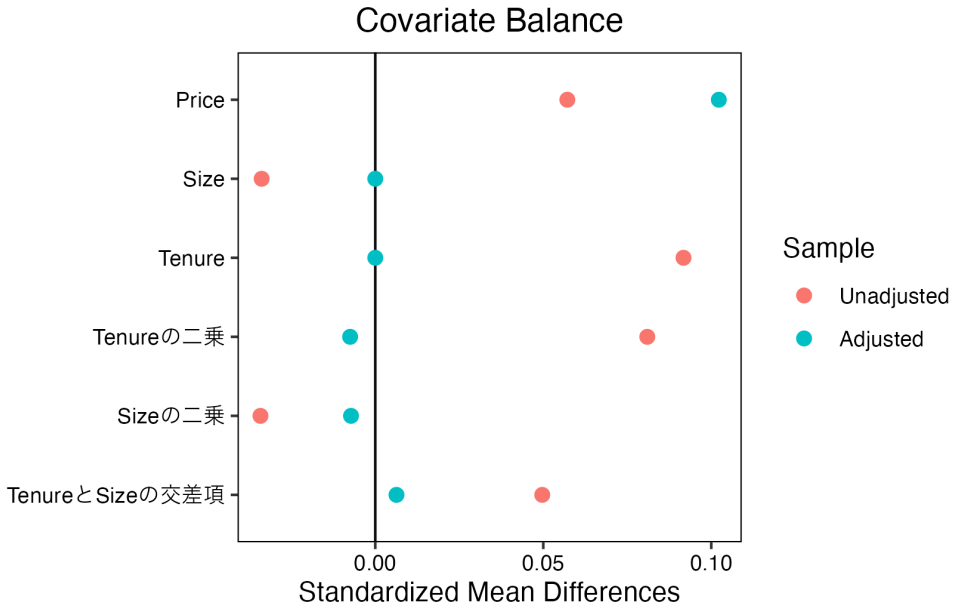
重回帰による推定は、D 間で X の平均値をバランスさせた上で、平均値を比較しています。また「母集団における OLS の結果」の推論に悪影響を与える、Weight の分散も可能な限り削減しています。

4.1.1　例

部屋の広さ (Size) と 築年数 (Tenure) をバランスさせた後に、2022($D = 1$)/2021($D = 0$) 年の平均取引価格差を推定します。$Price \sim D + Size + Tenure$ を OLS で推定したとします。

```
estimatr::lm_robust(
  Price ~ D + Size + Tenure,
  Data)
```

この推定によって得られる D の係数値は、以下のようなバランスを達成する Balancing Weights を用いた平均差と一致します。



赤点 (Unadjusted) は、バランス前の単純平均差を表します。価格が大きく上昇していますが、取引物件の部屋の広さは狭くなり、築年数は古くなっています。青点 (Adjusted)

は、OLS による暗黙のバランス後の差を示しています。Size や Tenure の平均値は完全にバランスしており、結果平均取引価格差も上昇しています。Tenure2 や Size2 は、築年数や部屋の広さの二乗項 (分散)、Tenure_Size は交差項 (共分散) を示しており、これらについてはバランスしていません。

4.2　OLS による分散や共分散のバランス

$Price \sim D + Size + Tenure$ を推定しても、Size や Tenure の平均値のみしかバランスできません。一見、これは OLS の致命的な弱点のように見えますが、簡単な修正によって解決できます

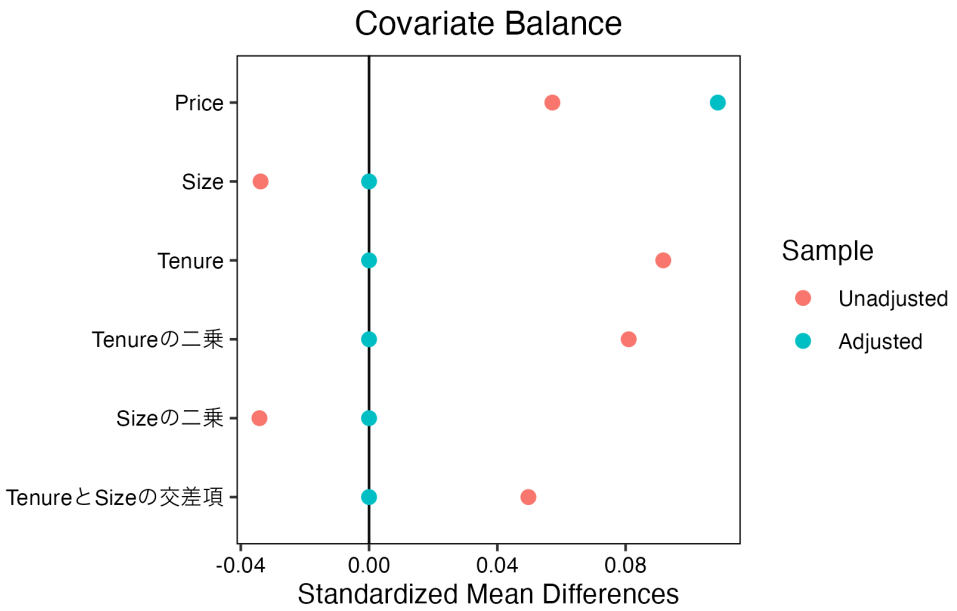
分散や共分散もバランスさせるためには、二乗項や交差項もモデルに導入したモデル

$$Price \sim D + Size + Tenure + Size^2 + Tenure^2 + Tenure \times Size$$

を OLS 推定します。

```
estimatr::lm_robust(
  Price ~ D + Size + Tenure +
    I(Size^2) + I(Tenure^2) +
    (Size + Tenure)**2,
  Data)
```

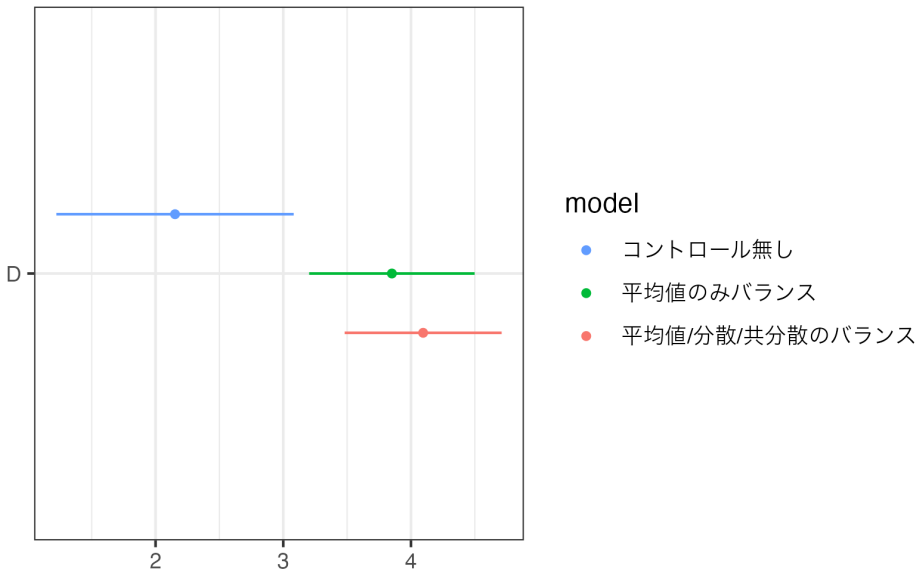
これによって、Size の二乗の” 平均値” などもバランスさせることができます。これは各変数の分散や共分散をバランスを意味します。結果、以下の図の通り、分散や共分散も Balance します。



4.3　母集団の推定方法

推定するパラメタ $(\beta_0, .., \beta_L)$ に比べて事例数が十分に大きければ、データ上の OLS の結果は、「母集団における OLS の結果」の優れた推定値です[*1]。特に近似的に計算される信頼区間を用いれば、母集団における OLS の結果を、定量的に議論できます。

以下では一切の X をバランスさせないケース、Size と Tenure の平均値のみをバランスさせるケース、平均に加えて分散と共分散もバランスさせたケースの推定結果を、95% 信頼区間とともに比較しています。



コントロールをしないケースに比べて、Size や Tenure をバランスさせると、2022 年と 2021 年の平均差が大きくなりました。一方で本データについては、平均値に加えて分散や共分散をバランスさせたとしても、あまり大きな変化は生じませんでした。

また信頼区間に焦点を当てるとバランス後の平均差の方は、前に比べて、より「狭く」なっています。これは、母集団における OLS の結果とデータ上の OLS の結果が、大きく乖離する可能性が減少している (と推測できる) ことを反映しています。

4.4　実践への示唆

実践における課題は、X について適切な複雑さを持つモデルを設定することです。しかしながら適切な複雑性は、母集団の特徴や事例数などに依存しており、一般な解決は困難で

[*1] 詳細に関心がある読者は、Chernozhukov et al. (2024) の 1 章などを参照ください。

す。ただし過去の実践の問題点、および実践上の示唆を得ることはできます。

多くの研究で X の二乗項や交差項を導入**しない**推定が行われてきました。しかしながらこのような推定は、平均値のみのバランスにとどまり、不十分な可能性が高いと考えられます。例えば、NBER Summer Institute 2018 における、Esther Duflo のチュートリアルでは、連続変数については二乗項、および全変数について交差項を導入した推定を行なっています。

複雑な推定は、母集団における推定結果の推論を困難にします。この問題は、事例数が少ない小規模データを用いた推定において深刻です。しかしながら現代的な分析環境のもとでは、1000 事例を超えるデータを用いた推定が一般的になっています。このため、バランスさせたい X の数が少ない場合、その二乗項や交差項を加えたとしても、悪影響は小さいと考えられます。少なくとも小規模事例を用いた推定よりも、より複雑なモデルを推定すべきであると考えられます。

バランスさせたい変数 X の数が多い場合、二乗項や交差項を導入するとモデルが爆発的に複雑化し、OLS では推定できなくなります。このような場合は、Esther Duflo のチュートリアル や Chapter 6 で議論する通り、機械学習を応用が有力です。

4.5 応用上の課題

OLS により暗黙のうちに計算される Weight は、平均値をバランスします。しかしながら、Balancing weights に求められる他の性質は必ずしも満たされません。

4.5.1 Researcher degrees of freedom

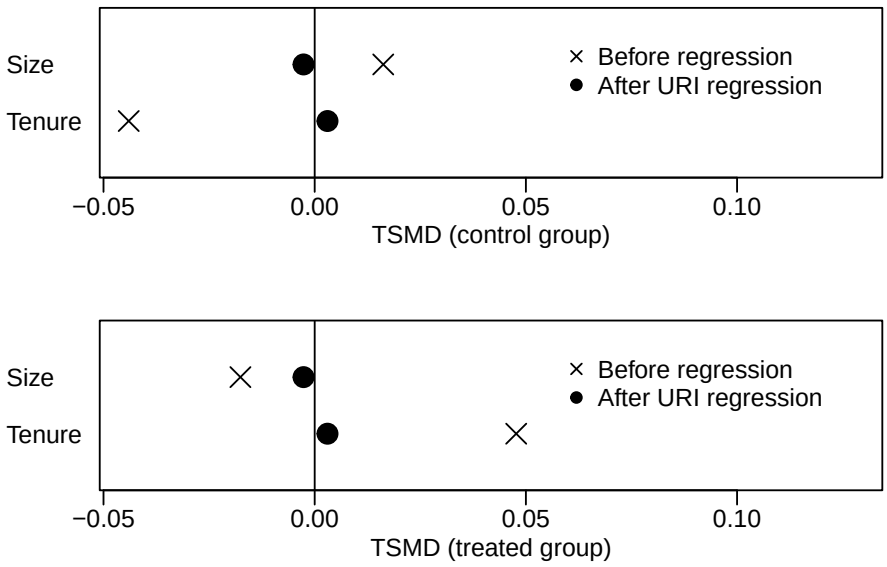
OLS においては、分布の特徴をどこまでバランスさせるのかが問題となります。事例数が十分あれば、3 乗項などの特徴もバランスさせることが可能です。しかしながら事例数が少ない場合、大量のモーメントをバランスさせると、推定誤差が大きくなってしまいます。このため十分な根拠を持った推定モデルの定式化が困難になります。

この問題に対して、Chapter 6 では機械学習を用いた改善方法を紹介します。

4.5.2 ターゲットの解釈の難しさ

バランス後の X の平均値がどのような水準になるのか、一般に不透明です。結果を解釈するためには、X の平均値は明確な水準、例えばデータ全体での平均値と一致させることが望まれます。しかしながら、OLS はそのような水準との一致を保証しません。

OLS によるバランス後の X の平均値について、lmw package により診断できます。



黒丸は OLS によるバランス後、ばつ印はバランス前の平均値を示しています。Control group は、$D = 0$ (2021 年)、Treatment group は、$D = 1$ (2022 年) の値です。0 線は、サンプル平均を示しています。

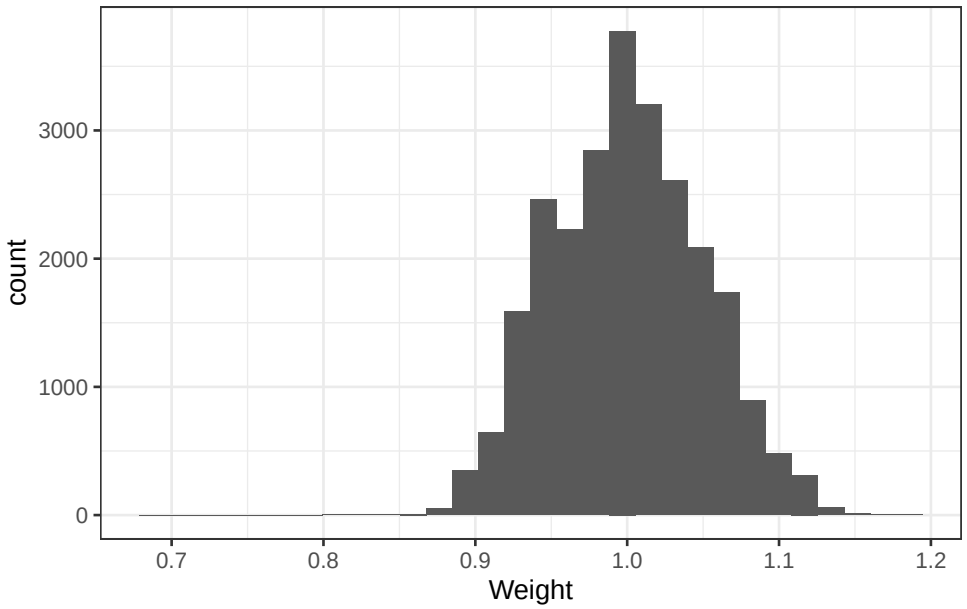
同図からバランス前は、2022 年については Size がサンプル平均よりも小さく、Tenure は長くなっています。黒丸を見ると OLS によるバランス後は、2022 年と 2021 年の間で平均差がなくなることが確認できます。ただし 0 線からは乖離しており、サンプル平均とは一致していないことが確認できます。

この問題の解決としては、サンプル平均にバランスさせることを明示的に要求した Balancing Weight の算出 (Chapter 5) が有力です。

4.5.3　負の荷重

Balancing weights は、正の値を取ることが望まれます。しかしながら OLS が生成する Weight は、負の値を取る可能性があり、ミスリードな推定結果をもたらす可能性があります。

lmw package は、OLS が生成する weights の値を計算します。例えば hist 関数により、ヒストグラムとして可視化できます。



本応用例では、負の weights は発生していないことが確認できました。

負の weights が発生しない方法としては、明示的な Balancing Weight の算出 (Chapter 5) や機械学習を活用した柔軟な推定 (Chapter 6) が有力です。

4.6 R による実践例

- D と X の交差項を含めたモデルの OLS 推定、およびその性質の診断は、以下のパッケージを用いて実装できます。
  - readr (tidyverse に同梱): データの読み込み
  - lmw: OLS が計算する balance weights を計算
    - Repository
  - estimatr: OLS を Robust standard error とともに計算
    - Repository
  - dotwhisker: 信頼区間の可視化
    - Repository

4.6.1　準備

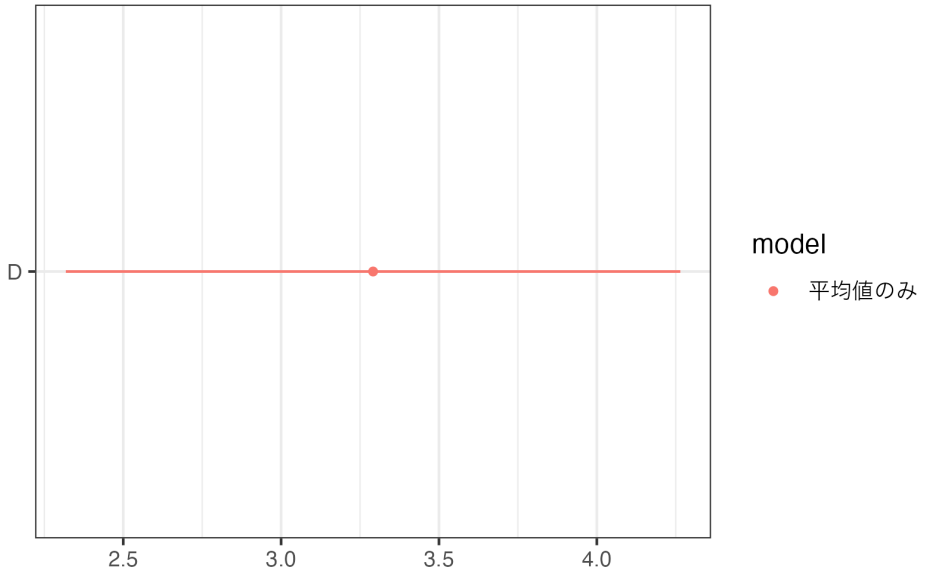
データを取得します。D として、取引年が 2022 か 2021 かで、1/0 となる変数を定義します。シンプルな比較分析について信頼区間は、データ分割は不要です。

```
Data = readr::read_csv("Public.csv") # データ読み込み

Data = dplyr::mutate(
  Data,
  D = dplyr::if_else(
    TradeYear == 2022,1,0
  ) # 2022 年に取引されれば 1、2021 年に取引されていれば 0
)
```

4.6.2　OLS によるバランス

D 間で Size,Tenure,StationDistance の平均値をバランスさせ、Price の平均値を比較します。

```
Model = estimatr::lm_robust(
  Price ~ D + Size + Tenure + StationDistance,
  Data) # OLS 推定

dotwhisker::dwplot(
  list(平均値のみ = Model),
  vars_order = "D") + # 信頼区間の可視化
  ggplot2::theme_bw() # 背景を白地化
```



D の係数値は 3.29 であり、20.06 ほど中心 6 区の物件の方が平均取引価格が高いことがわかります。

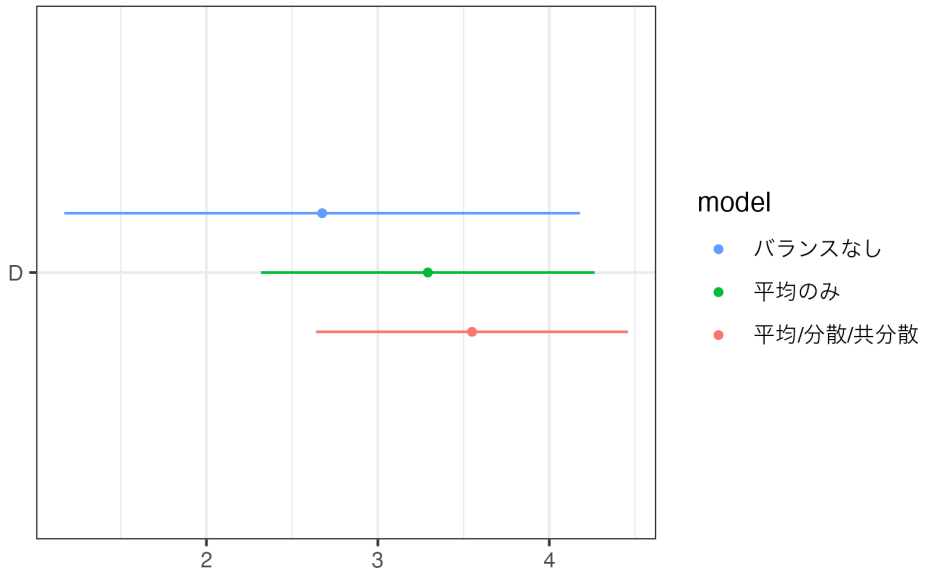
次に各変数の分散と共分散もバランスさせます

```
ModelLong = estimatr::lm_robust(
  Price ~ D +
    (Size + Tenure + StationDistance)**2 + # 交差項の作成
    I(Size^2) + I(Tenure^2) + I(StationDistance^2), # 分散
  Data)
```

バランスをしない単純比較も含めて、推定結果を比較すると以下のようになります。

```
ModelSimple = estimatr::lm_robust(
  Price ~ D,
  Data) # バランスなし

dotwhisker::dwplot(
  list(
    バランスなし = ModelSimple,
    平均のみ = Model,
    `平均/分散/共分散` = ModelLong
    ),
  vars_order = "D"
  ) +
```

```
  ggplot2::theme_bw()
```

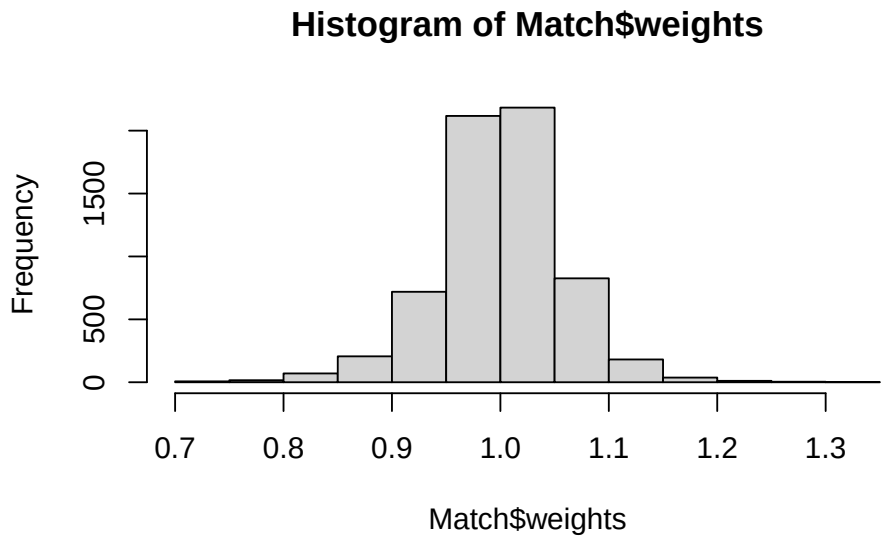


バランスすることで、推定値が大きくなり、信頼区間が縮小する (推定精度が改善する) ことが確認できます。

4.6.2.1　Balanced Weight

lmw パッケージの lmw 関数を用いれば、OLS が算出している Balance weights を計算できます。

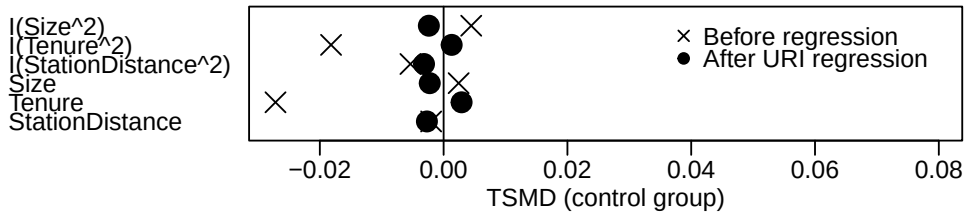
```
Match = lmw::lmw(
  ~ D + I(Size^2) + I(Tenure^2) + I(StationDistance^2) +
              (Size + Tenure + StationDistance)**2, # 平均、分散、共分散をバランス
  Data
) # Weight の算出

hist(Match$weights) # Weight のヒストグラムを算出
```

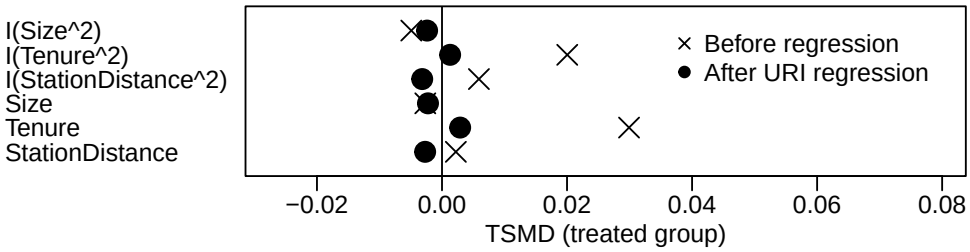


負の Weight が発生していないことが確認できます。

データ全体での平均値との乖離も、以下のとおり確認できます。

```
plot(summary(Match), abs = FALSE)
```





第 5 章

明示的な Balancing Weight の算出

Hainmueller (2012) や Zubizarreta (2015) では、Balancing Weight を明示的な” 最小化問題” の解として算出します。このようなアプローチは、満たすべき条件 (負の荷重が生じない/サンプル平均にバランスするなど）を課した上で、Balancing Weight が算出できます。このため OLS による暗黙の Balancing weight に比べて、より透明性の高い分析が可能です。

5.1 算出方法

Balancing Weight について、以下のような制約を貸します。

1. 全ての X_l について、その加重平均をデータ全体の平均値に一致させる:

$$\begin{aligned} &D = 1\text{における}(\omega(x,1) \times X_l)\text{の平均値} \\ =\ &D = 0\text{における}(\omega(x,1) \times X_l)\text{の平均値} \\ =\ &X_l\text{のデータ全体での平均値} \end{aligned}$$

2. 全ての Weight は非負の値をとる:

$$\omega(x,d) \geq 0$$

上記の制約を満たす $\omega(d,x)$ のなかで、最もばらつきが小さいものを Balancing Weight とします。ばらつきの測定方法は、いくつかの提案があります。

- Hainmueller (2012) : $\omega(x,d)$ の entropy divergence $\omega(x,d)\log(\omega(x,d)/q)$
  - q は base weight であり、例えば $q = 1/$事例数 が用いられる
- Zubizarreta (2015) : $\omega(x,d)$ の分散

特に entropy divergence を用いる Hainmueller (2012) の方法は、実際の計算も早く実用的です。以下では、実際の実装方法を紹介します。

5.2　R による実践例

- D と X の交差項を含めたモデルの OLS 推定、およびその性質の診断は、以下のパッケージを用いて実装できます。
  - readr (tidyverse に同梱): データの読み込み
  - WeightIt: entroy weight の計算
    - Repository
  - marginaleffects: WeightIt パッケージが計算する Weight を用いた推定
    - Repository

5.2.1　準備

データを取得します。D として、立地が中心 6 区か否かで、1/0 となる変数を定義します。

```
Data = readr::read_csv("Public.csv") # データ読み込み

Data = dplyr::mutate(
  Data,
  D = dplyr::if_else(
    LargeDistrict == "中心 6 区",1,0
  ) # 2022 年に取引されれば 1、2021 年に取引されていれば 0
)
```

ベンチマークとして、一切のバランスを行わない単純差、および OLS によるバランスを行った結果を示します。

```
            Estimate Std. Error  t value     Pr(>|t|) CI Lower CI Upper   DF
(Intercept) 38.03972  0.3182179 119.5399 0.000000e+00 37.41591 38.66354 6376
D           20.94057  1.2529064  16.7136 2.084367e-61 18.48446 23.39669 6376
```

```
                           Estimate  Std. Error    t value      Pr(>|t|)
(Intercept)             5.571162752 2.189748777   2.544202  1.097620e-
02
D                      19.124222089 0.663203786  28.836117 5.762617e-
172
```

```
Size                     0.940157758 0.097538720   9.638816  7.723573e-
22
Tenure                  -0.788224376 0.082744151  -9.526043  2.266689e-
21
StationDistance          1.067338032 0.151314422   7.053776  1.925476e-
12
I(Size^2)                0.007310055 0.001165973   6.269489  3.859676e-
10
I(Tenure^2)              0.013459763 0.001480646   9.090464  1.294755e-
19
I(StationDistance^2)    -0.010734191 0.009143823  -1.173928  2.404677e-
01
Size:Tenure             -0.015367241 0.001549637  -9.916673  5.169686e-
23
Size:StationDistance    -0.048089449 0.003899316 -12.332792  1.492403e-
34
Tenure:StationDistance   0.028634877 0.004083929   7.011599  2.599274e-
12
                            CI Lower     CI Upper   DF
(Intercept)              1.278517986  9.863807518 6367
D                       17.824119406 20.424324772 6367
Size                     0.748949031  1.131366486 6367
Tenure                  -0.950430768 -0.626017985 6367
StationDistance          0.770710827  1.363965238 6367
I(Size^2)                0.005024355  0.009595755 6367
I(Tenure^2)              0.010557197  0.016362328 6367
I(StationDistance^2)    -0.028659162  0.007190780 6367
Size:Tenure             -0.018405051 -0.012329431 6367
Size:StationDistance    -0.055733421 -0.040445478 6367
Tenure:StationDistance   0.020629000  0.036640753 6367
```

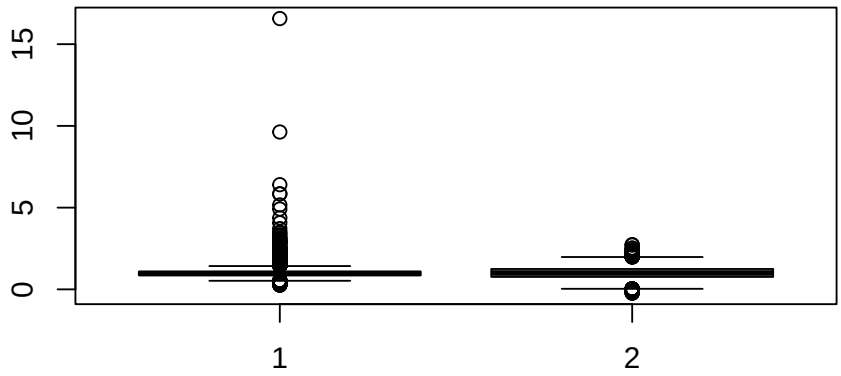
バランス前では中心 6 区の物件の方が、20.9 [18.5,23.4] 販売価格が平均的に高い傾向があります。この傾向は、部屋の広さや築年数、駅からの距離の平均値/分散/共分散をバランスさせても、大きくは変わりませんでした。

5.2.2 Entropy Weight によるバランス

D 間で Size,Tenure,StationDistance の平均値をバランスさせ、Price の平均値を比較します。まず WeightIt パッケージを用いて、Entropy Weight を計算します。また比較のた

めに lmw パッケージを用いて、OLS Weight も計算します。

```
Entropy = WeightIt::weightit(
  D ~ (Size + Tenure + StationDistance)**2 +
    I(Size^2) + I(Tenure^2) + I(StationDistance^2),  # 平均、分散、共分散をバランス
  data = Data,
  method = "ebal", # Entropy Weight を計算
  estimand = "ATE")

OLS = lmw::lmw(
  ~ D + I(Size^2) + I(Tenure^2) + I(StationDistance^2) +
    (Size + Tenure + StationDistance)**2,
  Data
)
boxplot(Entropy$weights,OLS$weights)
```



OLS を用いると負の Balancing weight が発生していることが確認できます。対して Entropy weight では、そのような weight は生じません。

```
OLS$weights |> summary()
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.2343  0.7615  1.0007  1.0000  1.2484  2.7393
```

```
Entropy$weights |> summary()
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2533  0.8622  0.9706  1.0000  1.0877 16.5677
```

Entropy Weight を用いたバランス後の平均差は、marginaleffects パッケージの avg_comparison 関数を用いて、以下のように計算できます。

```
WeightIt::lm_weightit(
  Price ~ D*(I(Size^2) + I(Tenure^2) + I(StationDistance^2) +
    (Size + Tenure + StationDistance)**2),
  data = Data,
  weightit = Entropy
  ) |>
  marginaleffects::avg_comparisons(variables = "D")
```

```
 Estimate Std. Error    z Pr(>|z|)     S 2.5 % 97.5 %
     19.9      0.625 31.8   <0.001 736.2  18.7   21.1

Term: D
Type:  probs
Comparison: 1 - 0
Columns: term, contrast, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high, predicted_l
```

引き続き 20 前後の平均価格差が算出されました。

第 6 章

機械学習の活用: 残差回帰

X の数が多い場合、OLS を用いた分布の特徴のバランスは、難しくなります。これはどのような特徴をバランスさせるのか、研究者の裁量の余地が大きくなりすぎるためです。このような研究者による推定式の定式化に起因する推定結果のブレは、Model Uncertainly (Varian 2014) とも呼ばれ、分析の信頼性を脅かす大きな要因となっています。

このような問題への対応として、機械学習を有効活用した、よりデータ主導のアプローチが注目されて来ました (Varian 2014; Urminsky, Hansen, and Chernozhukov 2019)。近年、機械学習の中でも教師付き学習の手法を、バランス後の比較に応用する方法について理解が急速に進みました。その中で教師付き学習 (より一般にはノンパラメトリック/データ適合的な推定) が持つ弱点である「収束の遅さ」[*1]を補完する方法が開発されています。このような方法を用いることで、信頼区間の近似計算など、母集団への含意がより明確な推定が可能となります。

本章は、OLS の直接的な拡張と解釈できる **残差回帰**への応用を紹介します[*2]。

6.1 OLS の一般化

OLS を用いて、X の分布を完全にバランスさせるには、以下のようなモデルを推定する必要がありまます。

$$Y \sim D + \underbrace{f(X_1, .., X_L)}_{X\text{についての極めて複雑な式}}$$

[*1] 「収束」の具体的な定義に関心がある方は、wikipedia の記事などを参照ください。

[*2] 機械学習は、より幅広い推定方法に応用できます。一般的な入門としては、Chernozhukov et al. (2024) を推奨します。他の入門資料としても、Wager (2024)、Schuler and Laan (2024)、Ichimura and Newey (2022)、Fisher and Kennedy (2021)、Hines et al. (2022)、Kennedy (2024) などの優れた教材が利用できます。簡潔な入門としては、Díaz (2020) がお勧めです。

$f(X_1, .., X_L)$ は、一般に無限個のパラメタを持つ式となります。

例えば 1 つの連続変数, X_1, のみのバランスが目標であるとします。この場合でも、推定式は以下のように無限個のパラメタを持つ式として定式化できます。

$$f(X_1) = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 ...}_{\text{無限和}}$$

このような無限個のパラメタを持つ式を、OLS で推定することは不可能です。あるいはパラメタの数が有限個であったとしても、データの事例数を超えると推定が不可能になります[*3]。

6.2　残差回帰

残差回帰は、$Y \sim D + f(X_1, .., X_L)$ が以下のように書き換えることができることに基づいています。

$$\underbrace{Y - (X\text{内での})Y\text{の平均値}}_{Y\text{の残差}} \sim \underbrace{D - (X\text{内での})D\text{の平均値}}_{D\text{の残差}}$$

すなわち Y の残差と D の残差を単回帰すれば、バランス後の平均値の比較を達成できます。

残る課題は、Y と D の平均値をデータから推定することです。このような推定は、X が多くある場合、困難な課題でした。以下、OLS 推定を再解釈することで、どのような問題が生じるのか示します。

6.2.1　OLS の再解釈

FWL 定理 (wiki) は、OLS も残差回帰として解釈できることを示しています。

OLS によって推定された D の係数値は、以下の手順で計算された値と厳密に一致します。

> **!** FWL(Frisch–Waugh–Lovell) 定理
>
> 0. Y/D/X を設定
> 1. 全データを用いて、Y , D の平均値の推定値 $f_Y(X), f_D(X)$ を $Y \sim \beta_0 + \beta_1 X_1 + .. + \beta_L X_L$ および $D \sim \beta_0 + \beta_1 X_1 + .. + \beta_L X_L$ を OLS 推定することで算出

[*3] パラメタと事例数、推定精度の詳細な関係性については、Chernozhukov et al. (2024) の 1 章とその引用文献を参照ください。

2. 残差 $Y - f_Y(X)$ と $D - f_D(X)$ を計算
3. 残差同士を回帰する: $Y - f_Y(X) \sim D - f_D(X)$

X が多い場合、Step 1. が大きな問題を抱えます。一般にパラメタ $(\beta_0..\beta_L)$ の数が、事例数に近い値になると、平均値の推定値は Y,D の実際の値に限りなく近づいていきます。このような現象は、過剰適合、あるいは過学習と呼ばれています。

機械学習の応用では、Step 1. を Y と D を X から予測する問題と捉えます。このような予測問題においては、OLS も含めたさまざまな推定手法を用いることができます。

6.2.2 教師付き学習の応用

ある一つの連続/カテゴリー変数 Y や D と (大量の) 変数 X の関係性の推定は、教師付き学習の中心的な研究課題です。中でも、(X 内での) Y や D の平均値の推定については、膨大な研究成果が蓄積されています[*4]。

一般に教師付き学習は、ある結果変数の予測を目的としています。予測の精度を平均二乗誤差で測定する場合、理論上最善の予測値は、(X 内での) 母平均であることが容易に証明できます。このため、(X 内での) 母平均を近似できるモデルの推定が、教師付き学習の実質的な目標となります。この性質を利用することで、既存の機械学習の方法で推定された予測モデルを、平均値を近似するモデルとして利用することができます。

予測モデルを推定する際には、Section 6.3 で議論する優れた推定上の性質を達成するために、交差推定を活用することが、一般に推奨されます (Naimi, Mishler, and Kennedy 2023)[*5]。

! 交差推定 (Cross Validation)

0. データを細かく分割 (第 1,..,10 サブグループなど)
1. 第 1 サブグループ**以外**で推定して、第 1 サブグループの予測値を算出
2. 第 2...サブグループについて、繰り返し、全事例に対して予測値を算出

推定に用いるアルゴリズムとしては、複数の予測モデルを組み合わせる Stacking 法が推奨されます (Ahrens et al. 2024; Naimi, Mishler, and Kennedy 2023)。これは多くの社会分析において、高い予測性能を期待できるアルゴリズムを事前に選択することが困難であるためです。

交差推定を活用した、残差回帰の具体的な推定手順は以下となります。

[*4] 詳細は、James et al. (2021) などを参照ください
[*5] 交差推定の利点/欠点については、Chen, Syrgkanis, and Austern (2022) などで重点的に検討されています。

機械学習を活用した残差回帰

0. Y/D/X および予測に使用するアルゴリズムを設定
1. 交差推定を用いて、Y と D を予測するモデル $f_Y(X)$ と $f_D(X)$ を推定する
2. 残差 $Y - f_Y(X)$ と $D - f_D(X)$ を計算する
   - 予測誤差とも解釈できます。
3. 残差同士を回帰する: $Y - f_Y(X) \sim D - f_D(X)$

6.3　母集団の推定方法

交差推定を併用した残差回帰の利点は、「母集団上で X を完全にバランスさせた後に計算した平均差」について、推論が可能なことです。特に交差推定と Stacking などの柔軟な方法で予測モデルを推定すれば、緩やかな仮定のみで信頼区間を近似計算が可能です (Chernozhukov et al. 2018)。

6.4　応用上の課題

機械学習を活用した残差回帰は、OLS の持つ幾つかの問題点を改善します。

十分な事例数があれば、「母集団上で X を完全にバランスさせた後に計算した平均差」を近似するため、負の荷重問題は生じません。

定式化そのものではなく、推定方法を選ぶため、Researcher degrees of freedom を適切に軽減できる可能性があります (Urminsky, Hansen, and Chernozhukov 2019; Ludwig, Mullainathan, and Spiess 2019) 。特に複数の推定方法を組み合わせる Stacking 法とシード値の適切な管理[*6]によっては、研究者への依存度を低下させられます。

問題点としては、ターゲットの解釈が容易ではないことが知られています。母集団上での残差回帰は、X を完全にバランスさせますが、そのターゲットは、以下となります。

$$\frac{(X\text{内での})D=1\text{の割合} \times (X\text{内での})D=0\text{の割合}}{(X\text{内での})D=1\text{の割合} \times (X\text{内での})D=0\text{の割合の平均値}}$$

このようなターゲットは、Overlap Weight とも呼ばれています。

Overlap Weight をターゲットとすると、D のばらつきが大きいグループの平均値を、最終的な比較結果に強く反映します。D のばらつきが無いグループ ($D=0$ または $D=1$ しか存在しない) の加重は 0 であり、最終的な比較結果に一才の影響を与えません。このため Overlap の仮定 (Important 1) を「必ず」満たすことできます。

[*6] Naimi, Yu, and Bodnar (2024), Schader et al. (2024); Zivich (2024)

このような優れた性質を持つ一方で、このターゲットをどのように解釈すればいいのか、一般に不透明です。特に X 間での D の偏りが大きい場合、Overlap Weight のばらつきも大きくなり、解釈が一層難しくなります。Zhou and Opacic (2022) など、Overlap Weight の解釈を提供する研究は行われていますが、確立された解釈は現状ありません。

6.5 R による実践例

- 以下のパッケージを使用
  - readr (tidyverse に同梱): データの読み込み
  - ddml: 残差回帰の実施
    - mannual

6.5.1 準備

ddml パッケージの関数を用いる場合、事前に Y と D はベクトルとして、X は行列として定義する必要があります。

```
set.seed(111) # シード値を固定

Data = readr::read_csv("Public.csv")

Data = dplyr::mutate(
  Data,
  D = dplyr::if_else(
    TradeYear == 2022,1,0
  ) # 2022 年に取引されれば 1、2021 年に取引されていれば 0
)

Y = Data$Price # Y の定義
D = Data$D # D の定義
X = Data |>
  select(
    District,
    Size,
    Tenure,
    StationDistance
  ) |>
```

```
  data.matrix() # X の定義
```

6.5.2　残差回帰

残差回帰は、ddml パッケージの ddm_plm 関数を用いて実装できます。

```
Model = ddml::ddml_plm(
  y = Y, # Y の指定
  D = D, # D の指定
  X = X, # X の指定
  learners = list(
    list(fun = ddml::ols), # OLS を使用
    list(fun = ddml::mdl_ranger) # RandomForest を使用
  ),
  shortstack = TRUE, # 簡略化した推定手順を指定
  silent = TRUE # Message を非表示
  )

Model |> summary()
```

```
PLM estimation results:


, , nnls

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.374      0.157   -2.38 1.73e-02
D_r            3.730      0.328   11.38 5.01e-30
```

summary 関数は、推定値 (Estimate)、標準誤差 (Std. Error)、t 値 (t value)、p 値 (Pr(>|t|)) を報告しています。バランス後の平均差は、D_r です。

95% 信頼区間は以下のように計算できます。

```
[1] 3.088172 4.372682
```

Reference

Ahrens, Achim, Christian B Hansen, Mark E Schaffer, and Thomas Wiemann. 2024. "Model Averaging and Double Machine Learning." *arXiv Preprint arXiv:2401.01645*.

Behaghel, Luc, Bruno Crépon, and Marc Gurgand. 2014. "Private and Public Provision of Counseling to Job Seekers: Evidence from a Large Controlled Experiment." *American Economic Journal: Applied Economics* 6 (4): 142–74.

Ben-Michael, Eli, Avi Feller, David A Hirshberg, and José R Zubizarreta. 2021. "The Balancing Act in Causal Inference." *arXiv Preprint arXiv:2110.14831*.

Bruns-Smith, David, Oliver Dukes, Avi Feller, and Elizabeth L Ogburn. 2023. "Augmented Balancing Weights as Linear Regression." *arXiv Preprint arXiv:2304.14545*.

Chattopadhyay, Ambarish, Christopher H Hase, and José R Zubizarreta. 2020. "Balancing Vs Modeling Approaches to Weighting in Practice." *Statistics in Medicine* 39 (24): 3227–54.

Chattopadhyay, Ambarish, and José R Zubizarreta. 2023. "On the Implied Weights of Linear Regression for Causal Inference." *Biometrika* 110 (3): 615–29.

Chattopadhyay, Ambarish, and José R. Zubizarreta. 2024. *Harvard Data Science Review* 6 (1).

Chen, Qizhao, Vasilis Syrgkanis, and Morgane Austern. 2022. "Debiased Machine Learning Without Sample-Splitting for Stable Estimators." *Advances in Neural Information Processing Systems* 35: 3096–3109.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* 21 (1): C1–68.

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. 2022. "Locally Robust Semiparametric Estimation." *Econometrica* 90 (4): 1501–35.

Chernozhukov, Victor, Christian Hansen, Nathan Kallus, Martin Spindler, and Vasilis Syrgkanis. 2024. "Applied Causal Inference Powered by ML and AI." *arXiv*

Preprint arXiv:2403.02467.

Chernozhukov, Victor, Whitney K Newey, and Rahul Singh. 2022a. "Automatic Debiased Machine Learning of Causal and Structural Effects." *Econometrica* 90 (3): 967–1027.

———. 2022b. "Debiased Machine Learning of Global and Local Parameters Using Regularized Riesz Representers." *The Econometrics Journal* 25 (3): 576–601.

Díaz, Iván. 2020. "Machine Learning in the Estimation of Causal Effects: Targeted Minimum Loss-Based Estimation and Double/Debiased Machine Learning." *Biostatistics* 21 (2): 353–58.

Fisher, Aaron, and Edward H Kennedy. 2021. "Visually Communicating and Teaching Intuition for Influence Functions." *The American Statistician* 75 (2): 162–72.

Gelman, Andrew, Jessica Hullman, and Lauren Kennedy. 2024. "Causal Quartets: Different Ways to Attain the Same Average Treatment Effect." *The American Statistician* 78 (3): 267–72.

Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis* 20 (1): 25–46.

Hines, Oliver, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. 2022. "Demystifying Statistical Learning Based on Efficient Influence Functions." *The American Statistician* 76 (3): 292–304.

Huling, Jared D, and Simon Mak. 2024. "Energy Balancing of Covariate Distributions." *Journal of Causal Inference* 12 (1): 20220029.

Ichimura, Hidehiko, and Whitney K Newey. 2022. "The Influence Function of Semiparametric Estimators." *Quantitative Economics* 13 (1): 29–61.

Imai, Kosuke, and Marc Ratkovic. 2014. "Covariate Balancing Propensity Score." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 76 (1): 243–63.

James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. 2021. *An Introduction to Statistical Learning*. Vol. 112. Springer.

Kallus, Nathan. 2023. "Treatment Effect Risk: Bounds and Inference." *Management Science* 69 (8): 4579–90.

Kennedy, Edward H. 2024. "Semiparametric Doubly Robust Targeted Double Machine Learning: A Review." *Handbook of Statistical Methods for Precision Medicine*, 207–36.

Ludwig, Jens, Sendhil Mullainathan, and Jann Spiess. 2019. "Augmenting Pre-Analysis Plans with Machine Learning." In *Aea Papers and Proceedings*, 109:71–76. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Naimi, Ashley I, Alan E Mishler, and Edward H Kennedy. 2023. "Challenges in Obtaining Valid Causal Effect Estimates with Machine Learning Algorithms."

American Journal of Epidemiology 192 (9): 1536–44.

Naimi, Ashley I, Ya-Hui Yu, and Lisa M Bodnar. 2024. "Pseudo-Random Number Generator Influences on Average Treatment Effect Estimates Obtained with Machine Learning." *Epidemiology* 35 (6): 779–86.

Opacic, Aleksei, Lai Wei, and Xiang Zhou. 2023. "Disparity Analysis: A Tale of Two Approaches."

Schader, Lindsey, Weishan Song, Russell Kempker, and David Benkeser. 2024. "Don't Let Your Analysis Go to Seed: On the Impact of Random Seed on Machine Learning-Based Causal Inference." *Epidemiology* 35 (6): 764–78.

Schuler, Alejandro, and Mark van der Laan. 2024. "Introduction to Modern Causal Inference." preparation.

Stuart, Elizabeth A, Gary King, Kosuke Imai, and Daniel Ho. 2011. "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference." *Journal of Statistical Software*.

Urminsky, Oleg, Christian Hansen, and Victor Chernozhukov. 2019. "The Double-Lasso Method for Principled Variable Selection."

Vafa, Keyon, Susan Athey, and David M Blei. 2024. "Estimating Wage Disparities Using Foundation Models." *arXiv Preprint arXiv:2409.09894*.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28 (2): 3–28.

Wager, Stefan. 2024. "Causal Inference: A Statistical Learning Approach." preparation.

Zhou, Xiang, and Aleksei Opacic. 2022. "Marginal Interventional Effects." *arXiv Preprint arXiv:2206.10717*.

Zivich, Paul N. 2024. "Commentary: The Seedy Side of Causal Effect Estimation with Machine Learning." *Epidemiology* 35 (6): 787–90.

Zubizarreta, José R. 2015. "Stable Weights That Balance Covariates for Estimation with Incomplete Outcome Data." *Journal of the American Statistical Association* 110 (511): 910–22.