母平均の「補助線」の推定 (ver 0.1.3)

川田恵介

2025-03-13

Table of contents

Preface

ある属性 X を持つ事例内での、変数 Y の平均値 (“条件つき” 平均) を推定する方法を紹介する、入門的なノートです[*1]。R の実例では、中古マンションの取引データを用いて、物件の属性 X (部屋の広さ、駅からの距離など) ごとに、平均取引価格 Y を推定します。

平均の推定値は、さまざまな実務で活用されています。中でも「Yの値を予測する」という課題において、中心的な役割を果たします。

伝統的には、OLS が母平均の推定方法として用いられてきました。近年では、OLS は母平均の仮想的な線型モデル (“補助線”) を推定する手法として、解釈できることが強調されています (Angrist and Pischke 2009; Aronow and Miller 2019) 。モデルの定式に誤りがあったとしても、推定結果は常に” 解釈” できることがその理由です。

また近年では、機械学習の手法も積極的に活用されています。本ノートでは、機械学習の手法を導入する動機として、OLS が” 研究者が設定した平均値のシンプルなモデル” を推定する方法であることを強調します。シンプルなモデルを推定する限りは優れた方法ですが、より複雑なモデルを推定したい場合はその有効性を失います。このような” 複雑なモデル” を推定する方法として、LASSO を紹介します[*2]。

[*1] より専門的な入門としては、Ding (2024) などを参照してください。
[*2] 他の予測モデルの推定方法については、James et al. (2021) 参照

第 1 章

要約の基本コンセプト

OLS や LASSO は、データが持つ特徴を**要約**するモデルを推定します。要約は、データ分析における中核的アイディアであり、その重要性の理解が分析の第一歩となります。まず本章では、要約の重要性を論じます。

1.1 観察できない変数が引き起こす問題

社会/市場分析における要約の必要性は、データから観察できない変数の存在にあります。データから観察できない変数の存在は、あらゆる事例分析の最も深刻な問題の一つです。このような変数への対処について、膨大な議論が蓄積されています。

観察できない変数がもたらす問題は、個別事例分析において特に顕著です。以下では、取引価格 (Price; 単位 = 100 万円) と物件の特徴を、事例分析から考察していきます例えば、2 億円で取引されている物件が、データの中に含まれていました。

この事例から、部屋の広さ (Size) が 105 平米で中心 6 区 (港、中央、千代田、新宿、渋谷、文京) に立地し、2 億円で取引された事例があることが確認できます。では、この事例をもとに、同じ属性を持つ物件も、2 億円で取引される傾向がある結論づけても良いでしょうか？ ほとんどの応用でこのような推論は、不適切です。

同じデータの中に、取引価格以外について、全く同じ特徴を持つ物件の取引事例が、以下の 3 件ありました。これらの事例と比較すると、2 億円はかなり高い価格での取引だったことがわかります。

なぜこのような取引価格のブレが生じるのでしょうか？ データの誤入力など潜在的な理

Price	Size	LargeDistrict
200	105	中心 6 区

Price	Size	LargeDistrict
200	105	中心 6 区
150	105	中心 6 区
92	105	中心 6 区
110	105	中心 6 区

由は複数ありますが、有力なのは**このデータに含まれない重要な変数**が存在することです。例えば、町丁目、最寄駅や公園の近くに立地するか否かなど、より詳細な属性も、価格決定において重要であると予想されますが、このデータには含まれていません。あるいは売り手や買い手の” 交渉力” を反映している可能性もあります。このように、多様な要因が取引価格に影響を与え、結果として取引価格の下振れ/上振れが生じます。

観察できない変数は不動産のみならず、個人や家計、企業、あるいは国レベルの分析でも同様の問題を引き起こします。観察できる変数 X が一致した事例内でも、観察できない変数は事例間で異る可能性が高く、結果 Y の値に大きな差が生まれます。

そして現実の社会や市場の複雑さを考慮すると、どれだけ詳細な調査を行ったとしても、Y に影響を与える全ての要因を観察することは困難です。

1.2　コンセプト: 集計

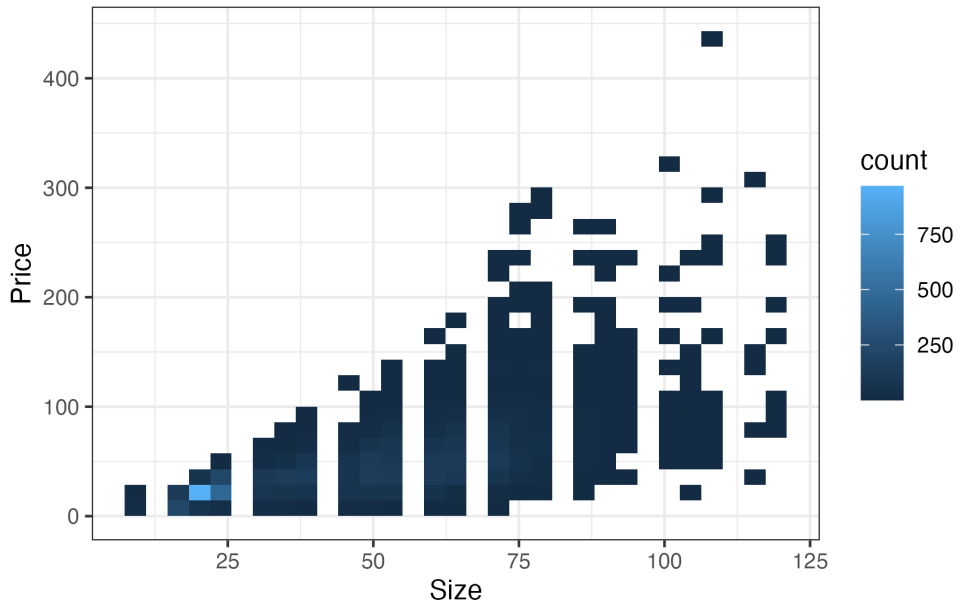
先の個別事例分析では、観察できない変数の偏りを確認する方法として、同じ X を持つ事例との整合性を確認しました。このようなアプローチの発展として、同じ X を持つ事例集団について、Y の特徴を要約する方法があります。例えば、平均値や分散、中央値、あるいは研究者による” 所見” や” 印象”、代表的な事例を紹介するなどです。

恣意的な分析を避けるためには、調査計画を立てる時点で、要約方法も決定し、分析を通じてコミットすることが重要です。このため分析内で、どのような「指標」を使用するのか、分析を開始する前に決定することが望まれます。

よく用いられる指標は、平均値です。

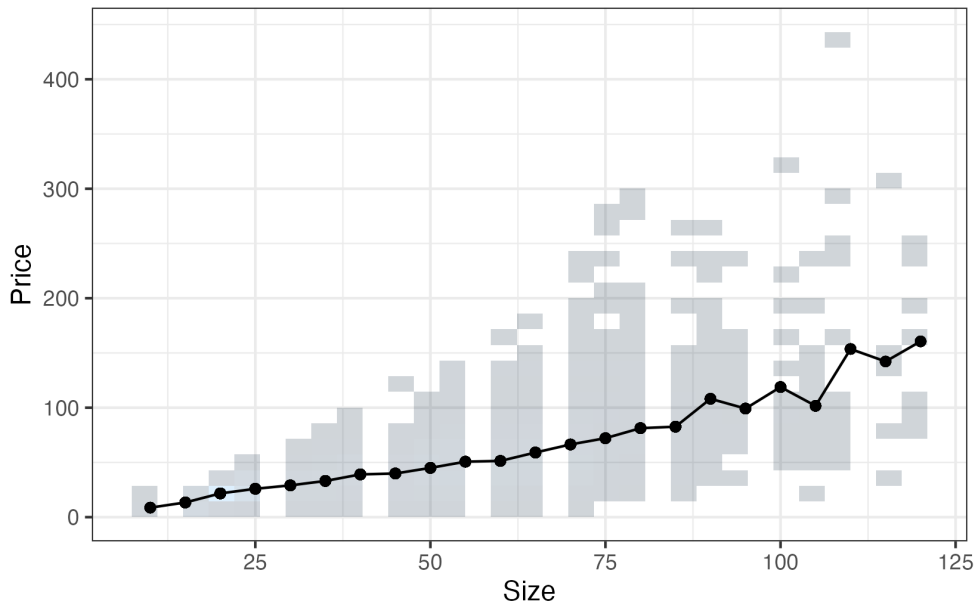
> ❗ データ上の平均値
>
> $$\frac{Y_1 + .. + Y_N}{N}$$
>
> ただし Y_i は第 i 事例の値、N は事例数を表す。

以下では、価格 (Price) と広さ (Size) について、データに含まれる事例の分布を Heat map で図示しています。



上記の散布図は、社会分析に用いるデータの持つ典型的な特徴を表しています。極めて乱雑であり、同じ X でも Y が異なる事例が多くなっています。これは、顕著な観察できない変数の影響を示唆しています。また X の値に応じた事例数の偏りも大きく、特に 100 平米を超える/20 平米を下回る物件の取引事例は少なくなっています。

以下の各点は、部屋の広さごとに計算された平均値を図示しています。



同図からは、部屋が広くなると取引価格は高くなる傾向が読み取れます。

多くの応用で、このような X ごとに集計するだけでは、現実の社会や市場の特徴について論じることは不適切です。本ノートでは、以下の少数事例の問題の解決に注力します。

1.3　少数事例の集計

平均値は有力な要約方法ですが、算出に使用する事例の数に注意してください。部屋の広さ (Size) ごとの事例数は、以下の通りです。

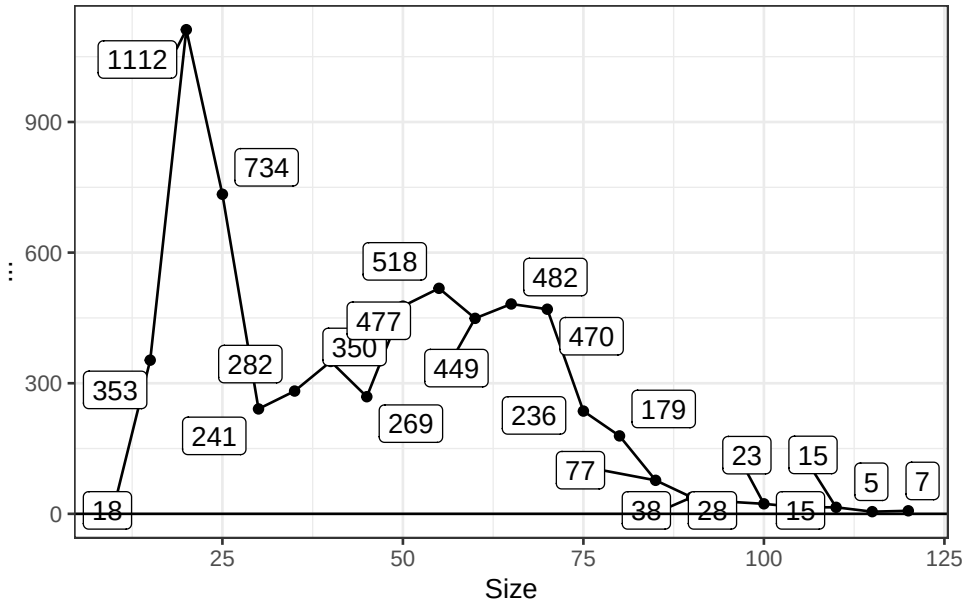


Figure1.1

特に 20 平米を下回る/100 平米を超える物件について、事例が少なくなっており、5 事例前後の組み合わせも散見されます。このような小規模な事例数からの計算は、多くの問題が発生します。そして、OLS や LASSO はそれに対応するための手法と解釈できることを強調します。

以下、Chapter 2 では、事例数が少ないと、平均値も各事例の観察できない変数の偏りの影響を強く受ける可能性を指摘します。Chapter 3 と Chapter 5 では、このような小規模事例の集計問題を緩和するための手法として、OLS や LASSO を紹介します。

第 2 章

母平均

2.1 頻度論

少数事例の集計が引き起こす問題を正確に理解するために、頻度論と呼ばれる概念的枠組みを導入します[*1]。

まず、「自分と同じ課題に取り組む他の研究者達」を想像してください。この研究者達は、あなたと同じ社会を対象に同じ手法を用いて分析しています。ただし共通のデータではなく、独立して収集した異なるデータを用いるとします。このような独立した研究者達は、あなたと同じ分析結果に到達するでしょうか？

簡単な理科の実験であれば、到達することは可能です。例えば水は、「水に不純物を入れない」、「大気圧が通常」などの実験の手続きを守れば、「誰がやっても」おおよそ 100 度で沸騰します。このため誰がやっても同じ得られる結果を、“科学的事実” として合意することができます。

対して現実社会における多くの現象について、独立した研究者が同じ結論に到達することは困難です。なぜならば、分析に用いる事例が異なるためです。ほとんどの応用で、独立して収集したデータが、研究者間で完全一致する可能性はほぼゼロです。

データに含まれる事例の偏りは、結論の不一致をもたらします。ある研究者には、公園に近い物件の取引事例ばかりが、“偶然” 集まってくるかもしれません。このような研究者のデータで計算された取引価格の平均値は、他の研究者と比べて、上振れる可能性が高くなります。このため要約した値であったとしても、研究者間で同じ値に合意できません。

人々が同じ結果を観察できない、という問題は深刻です。「独立した個人や組織が同じ結果を観察できるので推定結果を事実として認定する」、という強力な枠組みが活用できないからです。さらに言えば、他者の結果と乖離することが予想されるのであれば、自身が

[*1] 頻度論以外にはベイズ法と呼ばれる枠組みもあります。Lin (2024) などを参照ください。

分析した結果を” 信じる” 合理的な理由も無くなります。

この問題に対処するためには、何らかの概念的な枠組みが必要となります。

> ❗ 事例: 報道機関による世論調査
>
> 分析結果の不一致の典型例として、報道機関による世論調査が挙げられます。複数の機関による調査結果が、毎月公開されますが、その結果は各社で異なっています。理由は複数考えられますが、最も単純なものは、調査対象となる回答者が異なるためです。典型的な世論調査では、各社が独立して電話番号をランダムに発生させるなどの方法で、1000-2000 名ほどの回答者を極力ランダムに選んでいます。しかしながら、異なる調査に同じ人が回答する確率は、非常に低くく、必然的に調査結果も異なります。

2.2　コンセプト: 母集団とサンプリング

分析結果の不一致の問題を適切に論じるために、母集団とサンプリングという分析概念を導入します。これらの概念によって、すべての研究者に共通した**正答 (推定対象; Estimand)** と各々のデータから得られる**回答 (推定値; Estimates)** を、分離することを可能にします。

> ❗ 母集団とサンプリング
>
> 以下を想定する
>
> - 私たちが手にしているデータは、**母集団**という無数の事例の集団から、選ばれた (**サンプリングされた**) 事例から構成されている
>   - 母集団の全ての事例を用いて算出された値を**推定対象**と呼ぶ
>   - サンプリングされたデータから算出された値を**推定値**と呼ぶ
> - 本ノートでは、事例は完全ランダムに選ばる (ランダムサンプリング) を想定します。

すなわち、「私たちが得られる推定値は、ランダムに部分的な事例から得たものである」と想定します。一部の事例しか活用できていないため、母集団におけるすべての事例から計算された推定対象とは一致しません。

例えば日本における男女間家事分担格差の実態把握を行いたいとします。この場合、母集団は日本の家計全体となります。もし日本の家計全体における家事負担を把握できれば、推定対象は容易に回答可能です。しかしながら私たちのデータは、家計全体の一部であり、そこから得られる推定値（例えばデータ上の平均的な家事分担）は、推定対象 (日本全体の平均的な家事負担) とは異なります。このため日本全体では「女性の方が男性よりも家事負担が大きい」としても、データに含まれる事例が偶然偏り、「夫の家事負担がよ

り大きい」という推定値を誤って示す可能性があります。

母集団として、仮想的な集団を想定することもできます。例えば、あるコンビニのレジデータに、ある日の全ての来客者について、購入金額が全て記録されているとします。この場合、現実の来客者全てが記録されているため、母集団は存在しない、あるいは母集団の全てを観察できていると考えることも可能です。一方で、その日にコンビニを訪れた顧客は、潜在的な顧客の一部であると想定することもできます。皆さんも、よく利用するコンビニであったとしても、毎日は利用しないのではないでしょうか？ この場合、あなたは母集団である潜在的な顧客には属しますが、実際のデータには記録されない (偶然コンビニを利用しなかった) 可能性があります。

2.2.1 母平均

母集団に対しては、データの要約と同様の議論が適用できます。母集団においても、X が同じで合ったとしても、Y の値が異なることが想定されます。このため代表的な Y の値について、論じることが現実的です。

本ノートでは引き続き、平均値について論じていきます。母集団における平均値を以後、母平均と呼んでいきます。

> ❗ 母平均
>
> Y の母集団における平均値

データ上での平均値とは異なり、母平均を正確に知ることは不可能です。研究者は母集団を直接観察できないためです。ただしもし母集団が直接観察できるのであれば、すべての研究者は同じ母平均を計算するため、母平均の値について合意できます。

合意可能だが、直接計算できない母平均を、手元にある限られたデータから推測することが、本ノートの中心的な挑戦となります。

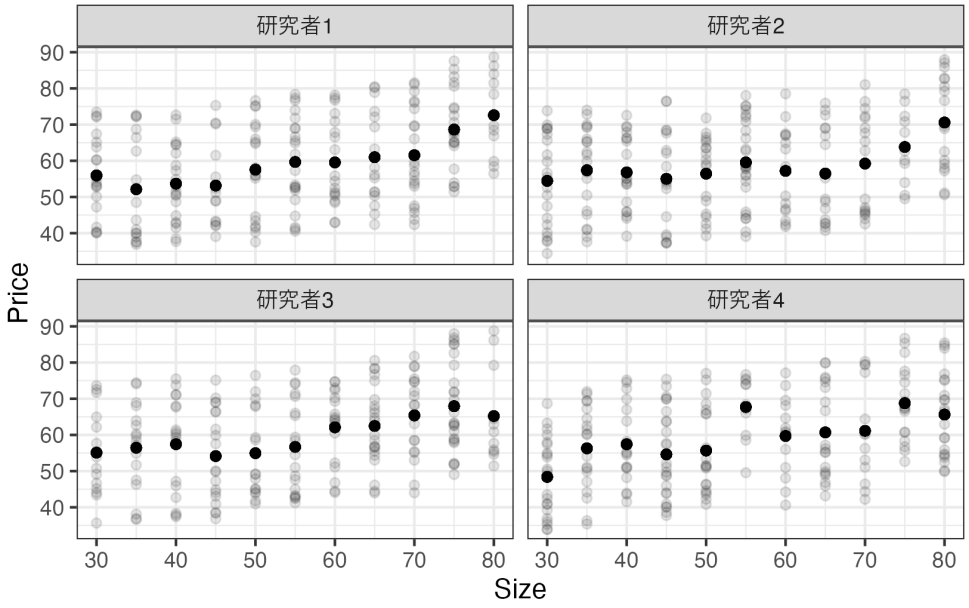
2.3 数値例

以上の概念を明確にするために、簡単な数値実験を行います。今、4 名の研究者が独立して 20 事例を集めたとします。各事例について、取引価格 Y と 部屋の広さ X がデータから観察できるとします

母分布は以下のように設定しています。

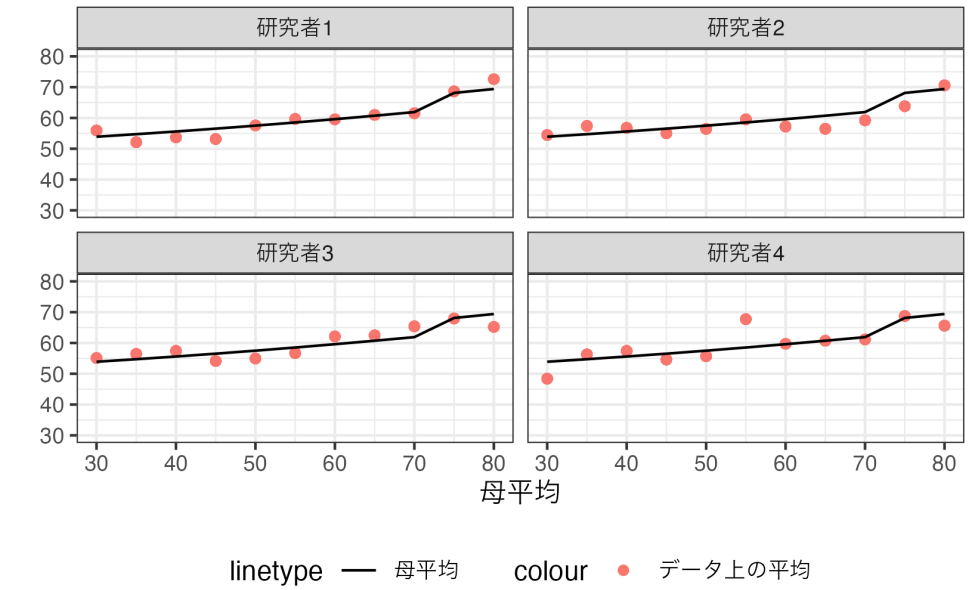
- 部屋の広さは、$X \in \{30, 35, 40, .., 80\}$ が同じ割合で存在
- 取引価格は、広さが 70-75 平米の間、急上昇する

以下の図は、4 名の研究者が手にするデータと平均値を図示しています。



平均値について、研究者間で大きな違いが見られます。

この図に母平均を上書きすると以下のようになります。ただし図を簡略化するために、各事例の値は排除します。



重要な点として、母平均とデータ上の平均は乖離していることを確認してください。また乖離の仕方は、研究者によって異なります。言い換えるならば、推定対象である母平均は

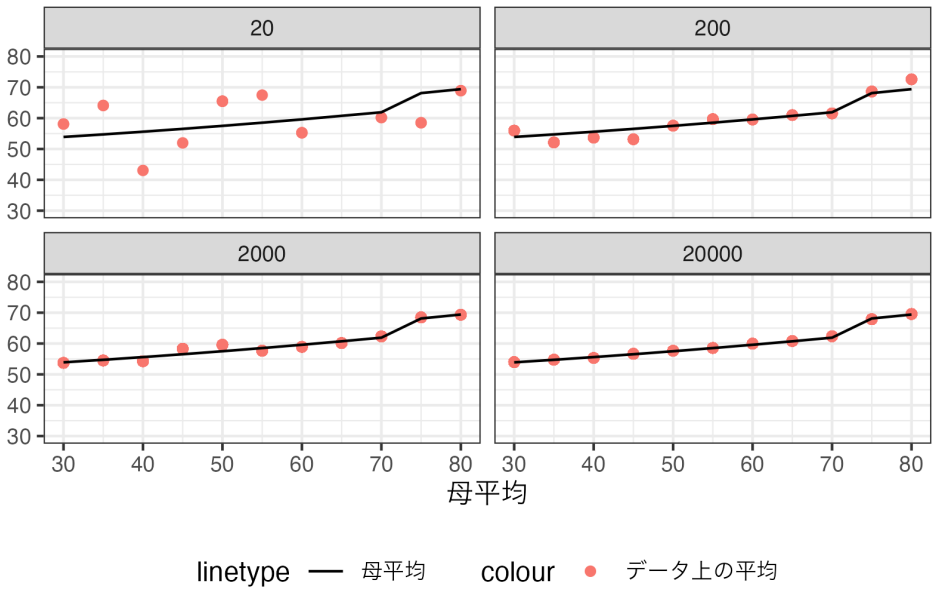
全員共通である一方で、推定値であるデータ上の平均値は異なっています。

2.4 大数の法則

母平均を知る方法は、各 X の組み合わせについて、ランダムサンプリングされた無限大の事例数で平均値を計算することです。これは大数の法則が成立し、データ上の平均値 (推定値) と母平均 (推定対象) が一致するためです。

> **!** 一致性
>
> 無限大の事例数をもつランダムサンプリングデータにおいて、計算された Y の平均値は、母平均と一致する。

以下では、事例数を 20,200,2000,20000 に随時変更した数値例を示しています。



20 事例では、母平均とデータ上の平均値が大きく乖離してますが、20000 事例ではほぼ一致していることが確認できます。

一致性は重要な理論的性質ですが、実践上の含意は限られています。多くの応用において、大量の X の組み合わせが発生する一方で、事例数は限られています。このため十分な事例数を確保することができず、小規模事例の集計の問題と向き合う必要があります。

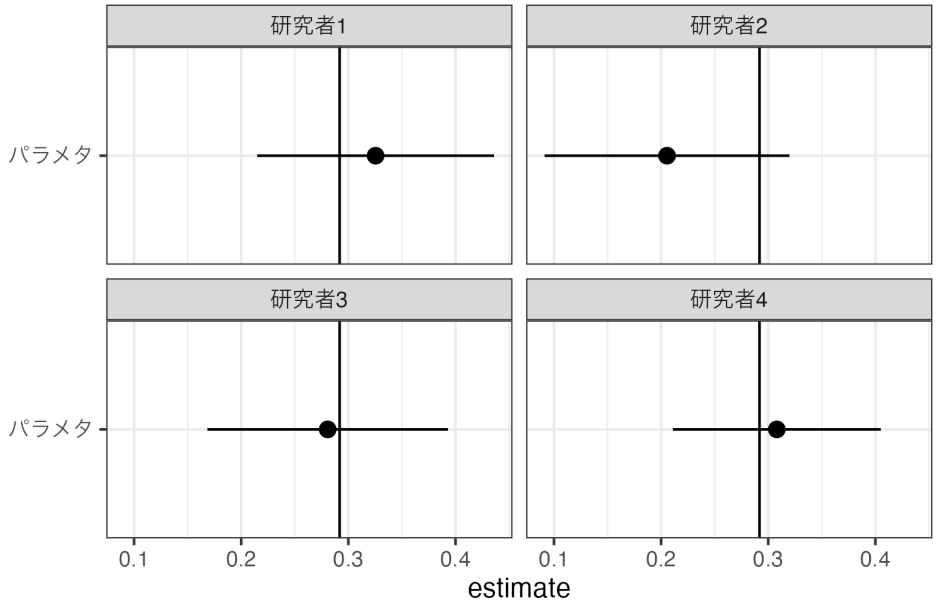
2.4.1　信頼区間

一致性は、データ上の平均値と母平均が一致を、無限大の事例数においてのみ保証します。現実の事例数では、データ上の平均値と母平均は一致せず、研究者間でのばらつきも残ってしまいます。

実際のデータ分析では、推定結果は**信頼区間**と共に示すことが一般的です。信頼区間とは、一定の確率[*2]で、母平均を含む区間です。

> ❗ 定義: 信頼区間
>
> 一定の確率 (初期値では 95%) で、母平均を含む区間

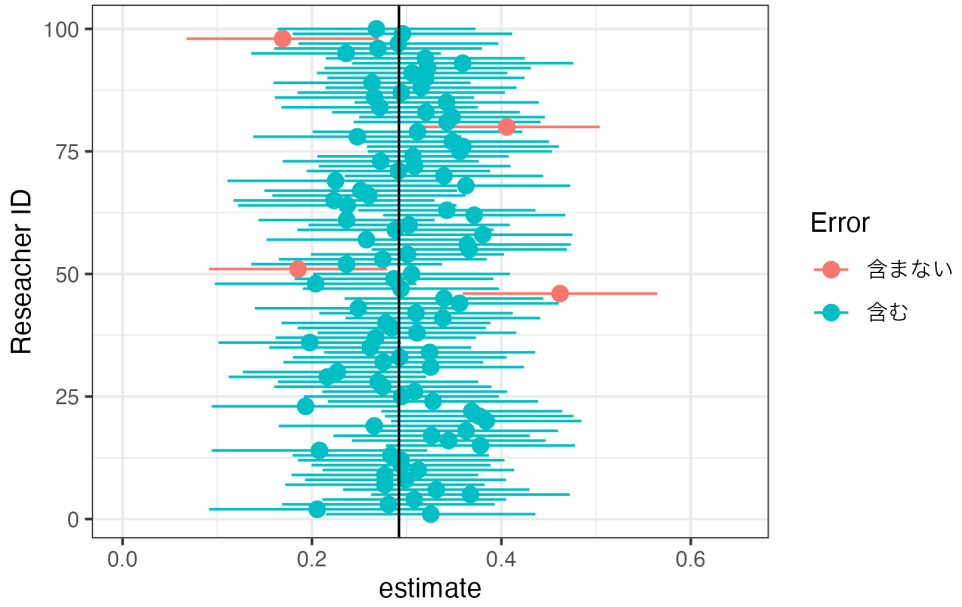
例えば 200 事例を用いて計算した平均値の信頼区間は以下です。



点で推定値、横線で 95 % 信頼区間、縦線が母平均を示しています。本数値例では、4 名の研究者全員が、母平均を含む信頼区間を獲得できています。

同じ数値例を、“100 名の研究者” 分実施した結果は以下です。

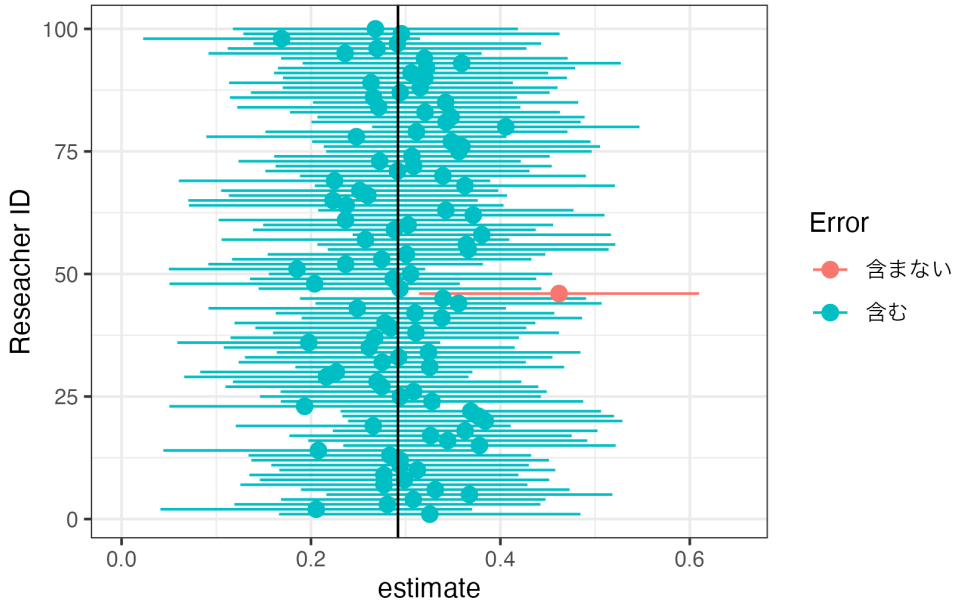
[*2] 多くの統計言語で、初期値では 95% が設定されています。



縦軸は研究者の名前 (ID) を示しています。赤字は、「不幸にも」母平均を含まない信頼区間が推定されてしまった研究者を示しています。100 名中 6 名 (概ね 95%) は、このような不幸に見舞われており、信頼区間の想定と整合的です。

実際の応用では、研究者自身は、自分が運の悪い 5% となるか、それとも幸運な 95% になるのかは分かりません。このため「5% のリスクは低い」として、自身が計算した信頼区間の中に真の値が” ほぼ” 含まれているとして、結論を論じます。

このリスクを変更することは容易です。たとえばリスクを 0.5% に変更した場合の 99.5% 信頼区間は以下です。



「不幸な」研究者は 100 名中 1 名のみとなりましたが、その代償として信頼区間が拡大しています。すなわち結論を不明確にする代わりに、ミスリードな主張をするリスクを減らしています。

「不幸な」研究者が発生する確率を 0% にすることは一般に不可能です。「100% 信頼区間」を計算しようとすると、区間の幅は無限大になります。

信頼区間を正確に推定することは、一般に極めて困難です。このため多くの実践で、近似的な信頼区間を算出し、報告します。この近似的な算出は、事例数が「十分にある」ことを前提としています。多くの数値実験から、具体的な事例数として 150 事例以上を要求されています[*3]。

部屋の広さと取引価格の例 (Figure 1.1) に戻ると、中間的な広さの物件については概ね 200 事例以上が確保されていますが、極端に小規模/大規模な物件については、事例数が 50 を大きく下回っています。このため信頼区間の計算が困難です。

[*3] 事例数と平均値、信頼区間の詳細な関係性については、Chernozhukov et al. (2024) の 1.2 章と 1.A 章などを参照ください

第 3 章

データ上での OLS

少数事例の要約を避けるためには、より” 大雑把な” 要約が必要となります。大雑把な要約の代表例として、線型モデルを紹介します。

線型モデルは、手元のデータから Y の平均値が持つ性質を簡便に捉えるモデルであり、現代のデータ分析でも頻繁に利用されます。

線型モデルを推定する方法としては、本章で最小二乗法 (OLS) および@sec-LASSO で LASSO を紹介します。

OLS による推定は、研究者によるモデルの単純化が求められます。適切な単純化がなされるのであれば、限られた事例数のもとでも、母平均の特徴を類推する有効な方法となり得ます (Chapter 4)。

3.1 線型モデル

Y の平均値と $X_1, .., X_L$ の関係性を記述するモデルを導入します。

❗ 線型モデル

$$Yの平均値 \simeq Yのモデル = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_L X_L$$

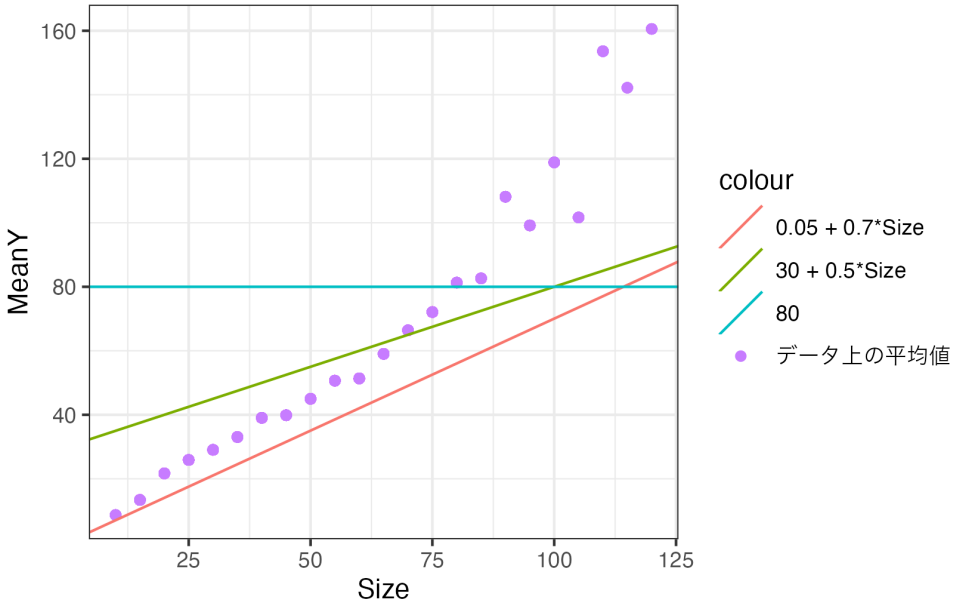
- $\beta_0, .., \beta_L$ は**パラメタ**と呼ぶ

以下では $\beta_0, .., \beta_L$ を決定する具体的な方法として、OLS を紹介します。

線型モデルをどのように解釈すれば良いでしょうか？ 最も実践的な解釈は、平均値の” 補助線” として捉えることです。

以下の図では、Price の平均値と Size の関係性を捉えるための 3 つの” 補助線” を書き込

みます。$\beta_0 + \beta_1 \times Size$ の、パラメタの値のみ変更しています。



データ上の平均値は紫の点で示しています。赤線は $0.05 + 0.7 \times Size$、緑線は $30 + 0.5 \times Size$ を示しています。

水色線は $\beta_0 = 80, \beta_1 = 0$ とした水平な「補助線」を示しています。

赤線と緑線は、平均取引価格が持つ「Size とともに上昇する傾向がある」特徴をある程度捉えています。対して水色線は、このような特徴を捉えられておらず、不適切であると考えれます。モデルの大枠が同じでも、パラメタ β の値によって、適切な要約か否かが決まってきます。

3.2　OLS

パラメタの値は、データに基づいて決定されることが通常です。代表的な決定方法としては、最小二乗法 (OLS) が挙げられます。

OLS の定義

1. 研究者が予測モデルの大枠を以下のように設定する

$$Y\text{のモデル} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_L X_L$$

2. 以下を最小化するように $\beta_0, .., \beta_L$ を決定する

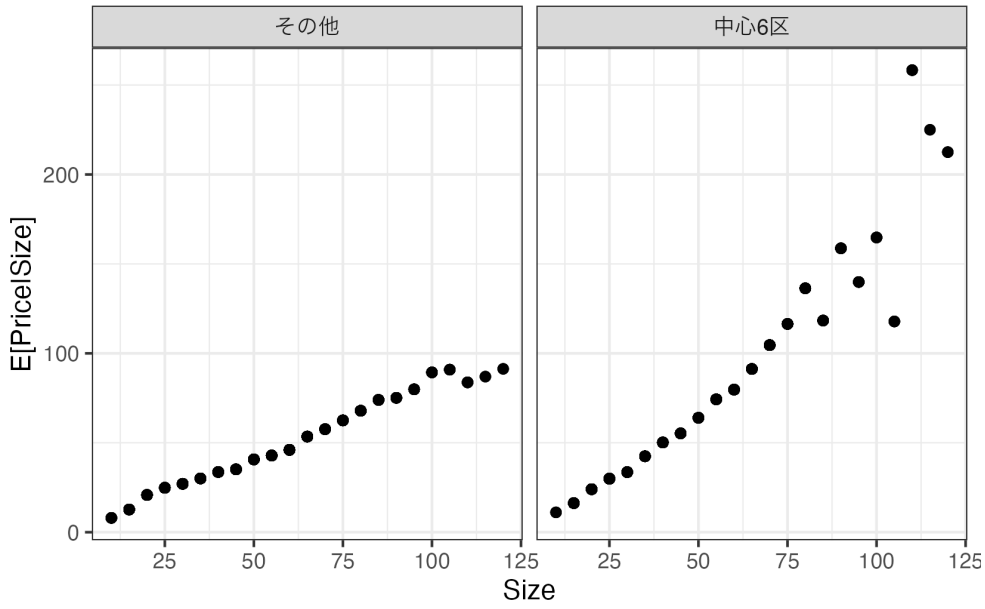
$$(Y - Y\text{のモデル})^2\text{のデータ上の平均}$$

OLS は、研究者が事前に大枠を設定したモデルを、データに最も適合するように推定する手法であると解釈できます。

3.3 実例

3.3.1 単回帰

２種類の X (Size と立地 (中心 6 区か否か)) について、取引価格の平均値を計算しました。



左側のパネルは中心 6 区、右側は他の区について、各 Size ごとに平均取引価格を計算しています。

平均値の最もシンプルな線型モデルとして、以下を推定してみます。

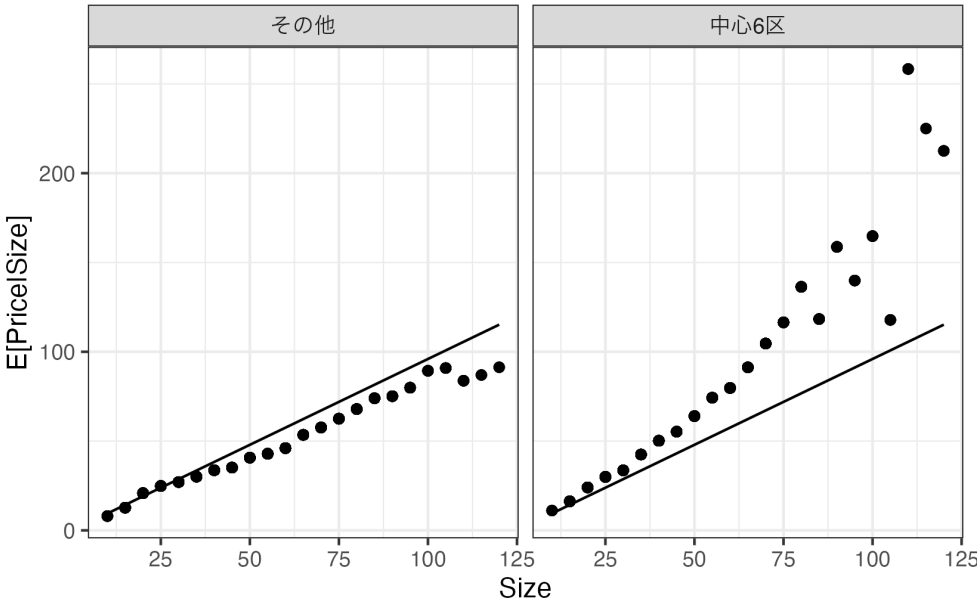
$$\text{モデル} = \beta_0 + \beta_1 \times Size$$

β_0, β_1 は、以下のデータ上の平均二乗誤差を最小化するように推定します。

$$(Y - \text{モデル})^2 \text{のデータ上の平均}$$

このような推定方法は、単回帰として教科書では紹介されてきました。

推定結果を図示すると、以下となります。



広い物件は取引価格が高くなる傾向を捉えることができています。しかしながら立地に関わらず同じモデルを当てはめており、中心６区の方が取引価格が高い傾向を捉えられていません。

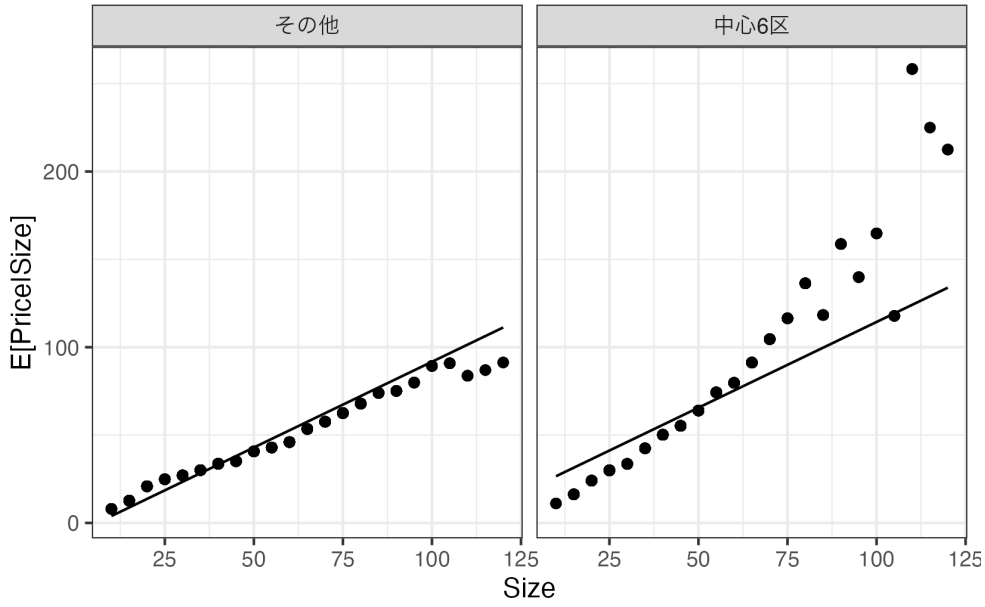
3.3.2　重回帰

立地と平均取引価格の関係性を捉えるために、以下のモデルの推定を試みます。

$$\text{モデル} = \beta_0 + \beta_1 \times Size + \beta_2 \times District$$

$District$ は、中心６区に立地していれば 1、それ以外では 0 を取ります。$\beta_0, .., \beta_2$ は引き続き、データへの適合度を最大化するように推定できます。このような推定方法は、重回帰として教科書では紹介されてきました。

推定結果を図示すると、以下となります。



中心 6 区の方が平均取引価格が高いという性質を上手く捉えています。しかしながら、中心 6 区において広い物件の取引価格が一段と上昇するという性質は捉えきれていません。

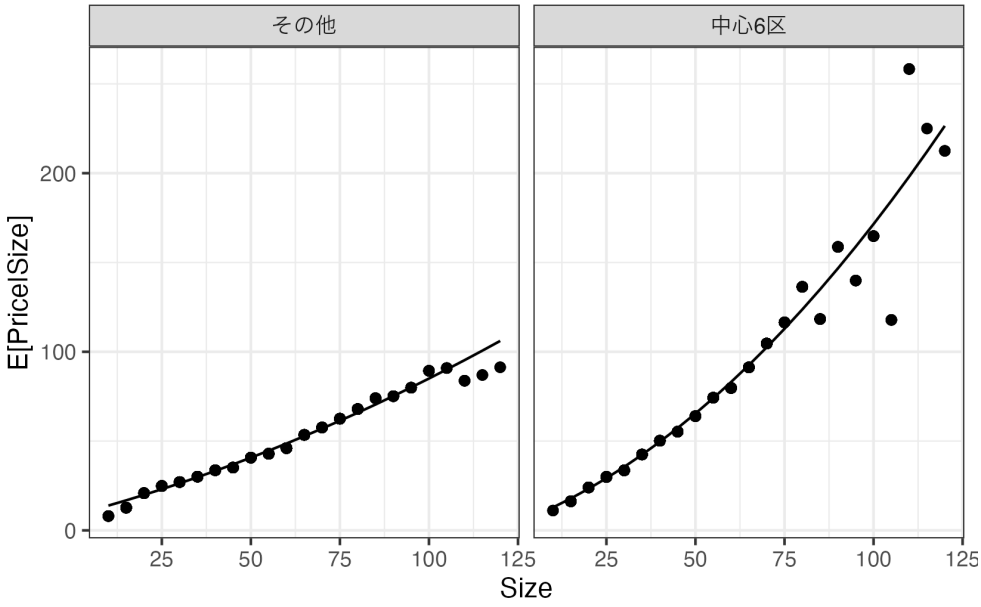
3.3.3 交差項と高次項の導入

母平均が持つ複雑な性質を捉えるために、交差効果と高次項を導入し、さらに複雑なモデルを推定してみます。

$$\text{モデル} = \beta_0 + \beta_1 Size + \beta_7 District + \underbrace{\beta_2 Size^2 + .. + \beta_6 Size^6}_{\text{高次項}}$$

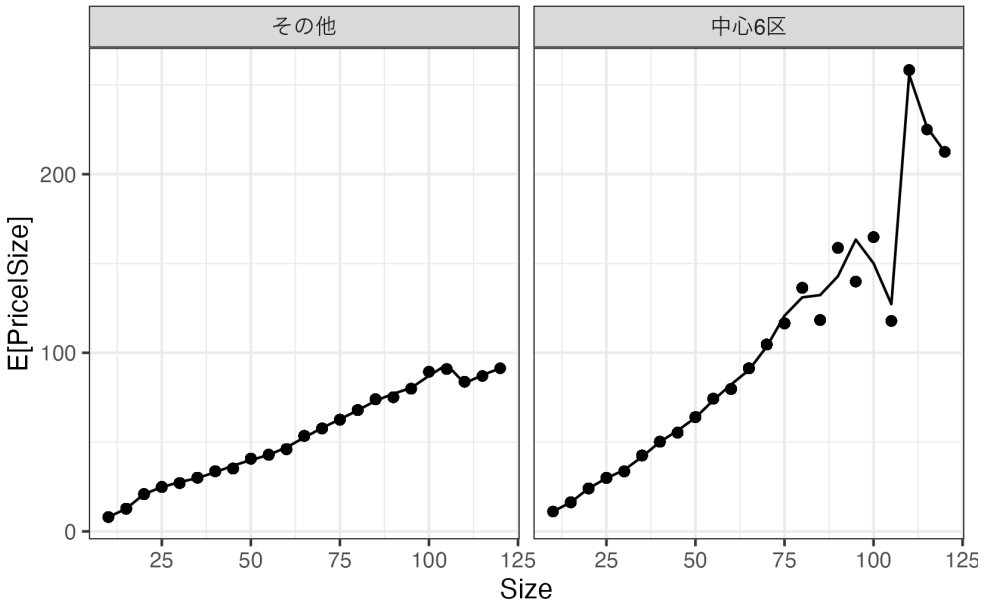
$$+ \underbrace{\beta_8 Size \times District + .. + \beta_{14} Size^6 \times District}_{\text{交差効果}}$$

このような複雑なモデルであったとしても、データへの適合度を最大化するように推定できます。



3.3.4　複雑なモデルの弊害

より複雑なモデルを最小二乗法で推定すると、データへの適合度が改善し、モデルをデータ上の平均値により近づけることができます。例えば、以下の図では Size の 10 乗まで加えた推定を行なっています。



このモデルでは、特に中心 6 区外に立地する物件について、ほぼほぼデータ上の平均値を近似するモデルが推定されています。さらにモデルを複雑化すると、データ上の平均値を”

なぞる” モデルが推定されます。

しかしながら、母集団の特徴を捉えることを目標とするのであれば、このことは必ずしも望ましいとはいえません。いうまでもなく、平均値をなぞるモデルは、単なる平均値とよく似た性質を持ちます。このため、Chapter 1 で議論した少数事例の集計の問題を引き起こしてしまいます。

以上の問題は、過剰適合/過学習の問題と呼ばれています。

> ❗ 過剰適合/過学習
>
> 複雑なモデルを、少ない事例数で推定した結果、データへの当てはまりは高くなるが、母平均からは乖離する現象

3.4 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

データを取得します。

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

lm 関数を用いて OLS を推定します。

```
OLS = lm(Price ~ Size + Tenure + StationDistance, # Y ~ X
         Data # 使用するデータの指定
         ) # OLS

OLS
```

```

Call:
lm(formula = Price ~ Size + Tenure + StationDistance, data = Data)

Coefficients:
    (Intercept)             Size           Tenure  StationDistance
        19.7206           1.0199          -0.6392          -1.3851
```

Coefficients が β の推定値を示しています。例えば、推定された線型モデルにおいて、Size(部屋の広さ) と平均取引価格は正の関係性がありますが、Tenure(築年数) と StationDistance(駅からの距離) は負の関係性が見られます。

第 4 章

母集団上での OLS

Chapter 2 では、データ上の平均値と母平均との関係性について論じました。データ上で実施した OLS は、平均値と同様に、母集団の特徴ついて何らかの含意を持つでしょうか？

伝統的な教科書では、研究者が設定した線型モデルが、母平均の妥当なモデルであれば、OLS により推定された線型モデルは、母平均の優れた推定値であることが強調されます。しかしながら社会科学のほぼ全ての応用において、研究者が設定するモデルは誤定式化 (misspecification) を犯しており、どのように推定しても母平均から乖離していると想定すべきです。

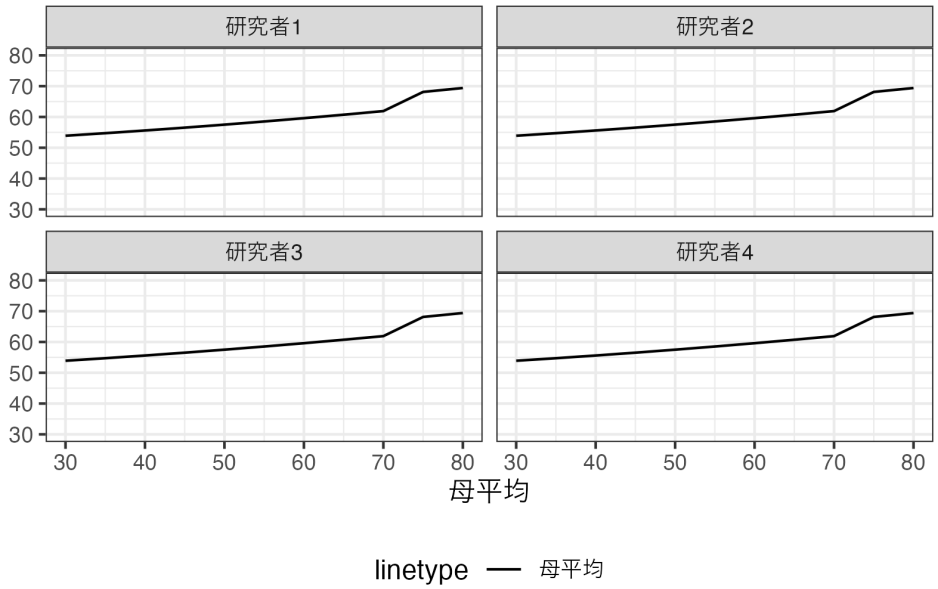
本章では、誤定式化を犯しているケースにおいても、OLS は母集団の特徴について明確な示唆を持つこと (Angrist and Pischke 2009; Aronow and Miller 2019) を紹介します。また誤定式化を減らすためには、モデルを複雑化する必要がありますが、パラメタ推定の「精度」との間にトレードオフが生じることも強調します。

このことが持つ示唆は重要です。経済学などの社会科学における実証研究は、しばしば推定するモデルが単純すぎるという批判を受けてきました。これは的を射た批判であり、現実社会の複雑さに比べると、如何なるモデルも単純すぎると想定すべきです。しかしながら、モデルの推定に用いることができる事例数に限りがあることも、同時に考慮する必要があります。現実に合わした極めて複雑なモデルを推定しようとすると、そのパラメタの推定精度が大幅に低下してしまいます。限られたデータと現実の複雑性の間で、適切な落とし所を見つけることが重要となります。

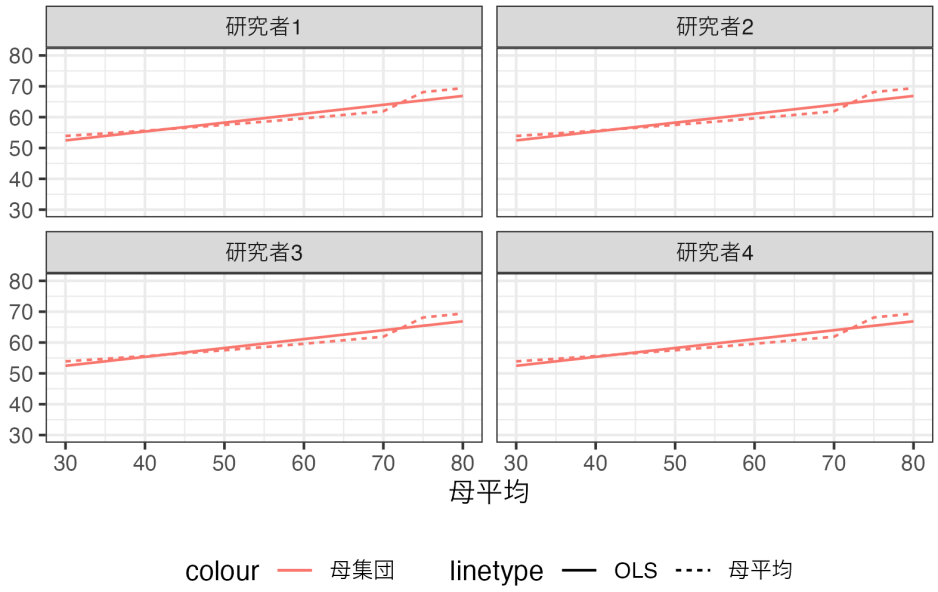
4.1 Population OLS

OLS の推定結果は、母集団上での**仮想的な** OLS (Population OLS) の結果を推定していると解釈することができます。少しわかりにくい考え方なので、Chapter 2 における数値例とともに確認します。

仮想的な 4 名の研究者が、同じ母集団の特徴を解明しようとしているとします。母平均は、全研究者共通で、以下の通りとなります。



もし母集団上で、モデル $f(X) = \beta_0 + \beta_1 X$ を OLS 推定できれば、以下の推定結果を得ることができます。



Population OLS の結果は実線、母平均は点線で表しています。以下の要点に特に注意してください。

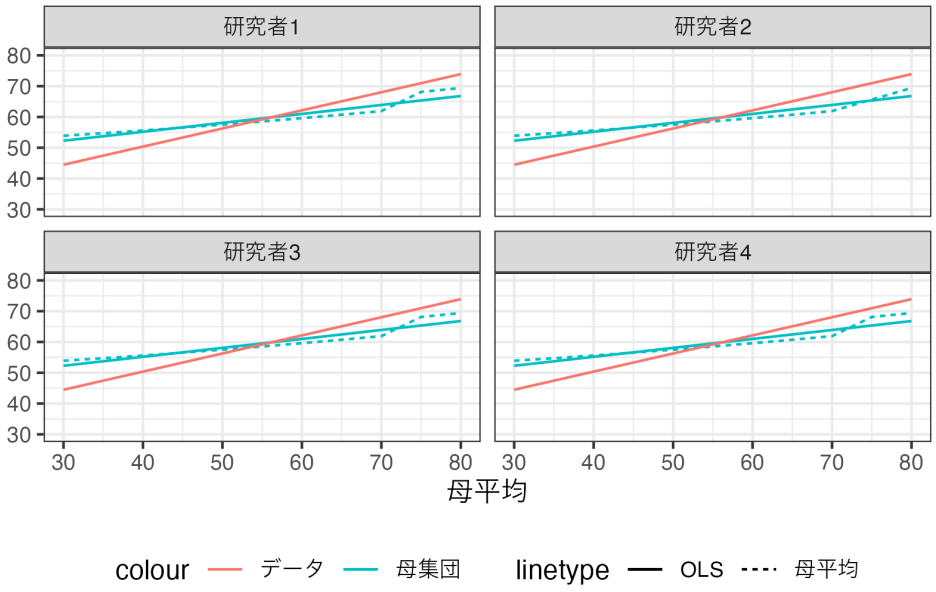
- Population OLS は、同じ” データ”(母集団) を用いて推定しているので、全ての研究者が同じ推定結果となります。
- 母平均とは必ずしも一致しません。本例において、母平均は 70~75 平米にかけて、平均取引価格が急上昇しています。しかしながら、“一直線” のモデルを推定しているため、このような傾向はモデルに反映されません。

二点目は誤定式化の問題と呼ばれ、近年大きな議論がなされてきました。

> **!** 誤定式化の定義
>
> ある線型モデル $\beta_0 + \beta_1 X_1 + .. + \beta_L X_L$ について、($X_1..X_L$ における)Yの母平均とモデルを一致させるような $\beta_0..\beta_L$ は存在しない

4.2 OLS = Population OLS の推定

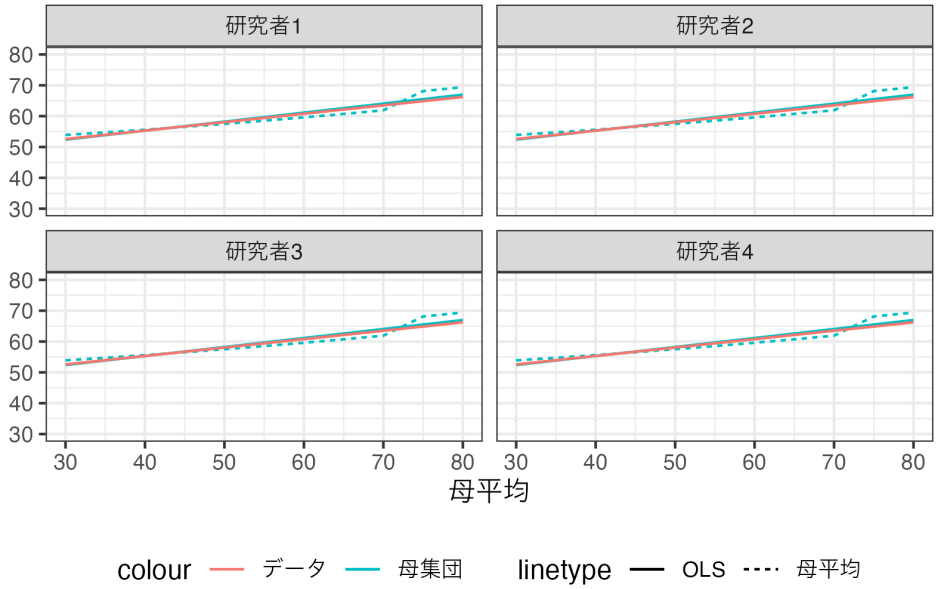
現実において実行可能な推定は、母集団ではなく、そこからランダムに選ばれた事例を用いた OLS です。ここでは 50 事例を収集したとします。各研究者は独自にデータ収集を行うため、データ上で行う OLS の結果には違いが生じることに注意してください。



母平均は青の実線、Population OLS は青の点線、データ上での OLS の結果は赤の実線で示しています。

データ上の OLS 推定から得られるモデルは、母集団上での OLS と一致はしていませんが、かなり近い性質を持っています。事例数を増やすとさらに近くなることも確認できま

す。以下では 5000 事例まで増やしています。



全ての研究者について、Population OLS とデータ上での OLS の乖離 (赤と青の実践の乖離) は、“目視” できないほど小さくなっています。

Population OLS とデータ上での OLS は、母平均とデータ上での平均値と類似した理論的関係性を持ちます。

> ❗ 性質: 大表本性質
>
> - 一致性: 事例数が無限大に大きくなると、データ上での OLS は Population OLS の結果と一致する。
> - 信頼区間: 事例数がパラメタの数に比べて十分多いと、Population OLS の結果についての信頼区間を近似計算できる。

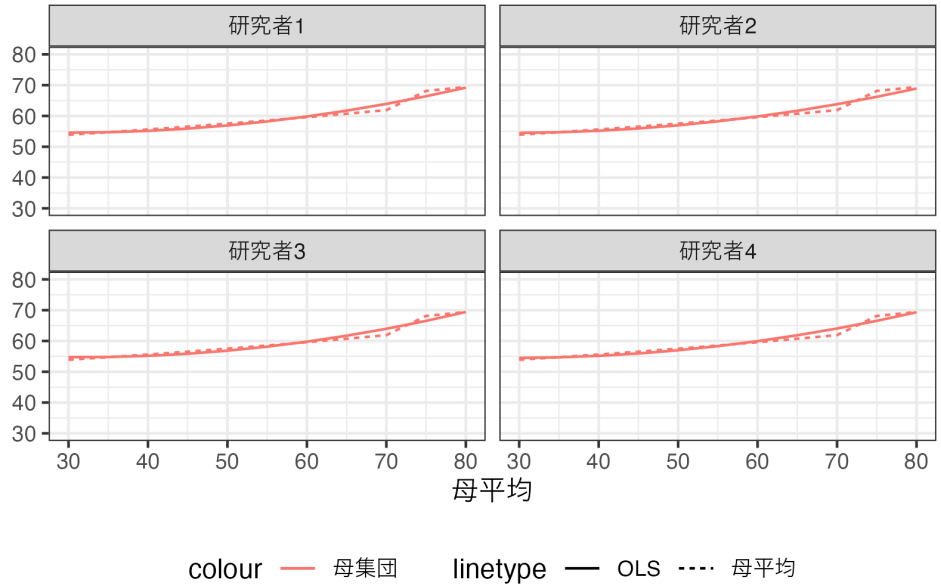
上記の性質はあくまで、データ上での OLS と Population OLS との関係性を論じていることに、改めて注意してください。大標本理論は、データ上での OLS を Population OLS の推定値とみなすことを正当化します。

しかしながら、母平均の推定値とみなすためには、誤定式化がないことが前提となります。なぜならば**誤定式化が存在すると**、Population OLS と母平均は一致しないので、データ上での OLS も母平均と一致することはあり得ません。

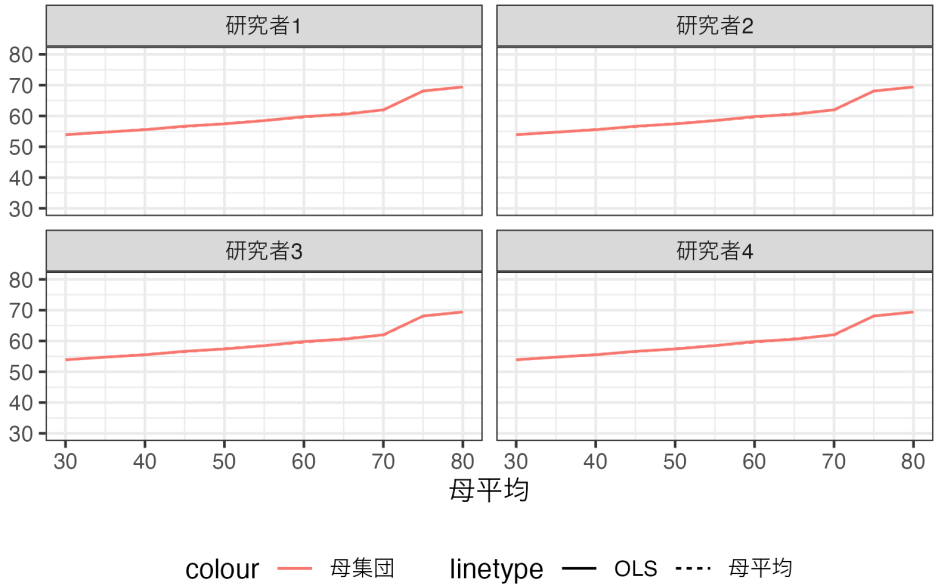
4.3　モデルの複雑化

実際の応用では、あらゆるモデルは誤定式化を犯している、少なくともその可能性を排除できない、と想定すべきです。特に社会現象や個人行動の大部分は Black Box であり、信頼できるモデルを想定することは、事実上不可能です。ただし誤定式化による母平均と Population OLS の乖離を削減することは容易です。しかしながら、その代償として、Population OLS とデータ上での OLS の乖離が大きくなる可能性があります。

モデル $f(X) = \beta_0 + \beta_1 X$ に変わって、$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2$ を推定してみます。仮想的な Population OLS の結果は以下です。



$f(X) = \beta_0 + \beta_1 X$ の Population OLS の結果よりも、母平均に近づいていることが確認できます。さらに X の 8 乗まで加えると ($f(X) = \beta_0 + \beta_1 X + .. + \beta_8 X^8$)、Population OLS は母平均をほぼ近似できることが確認できます。

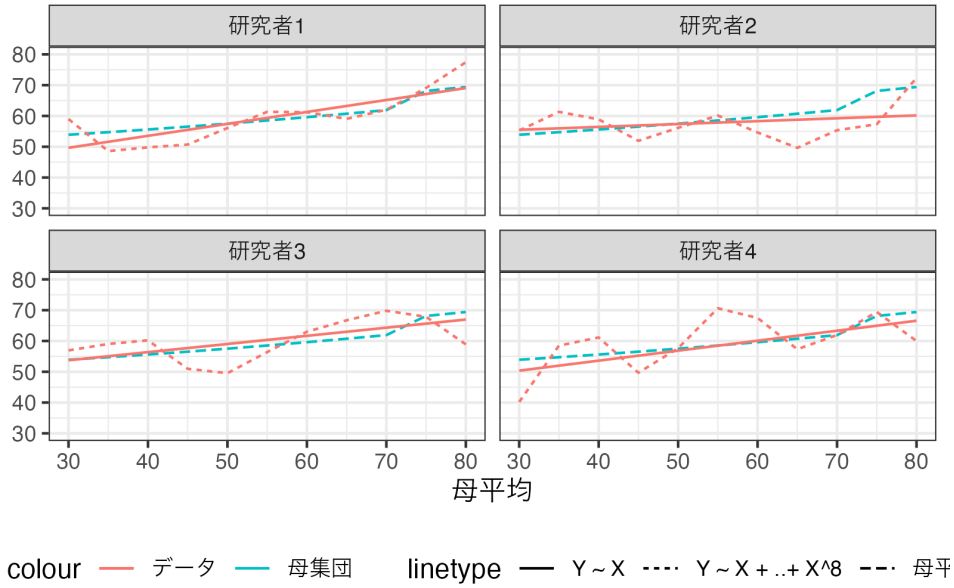


一般にモデルを複雑化する (β の数を増やす) と、Population OLS と母平均は**必ず**近づきます[*1]。

もし母平均の推定が目的なのであれば、極力複雑なモデルを推定すべきでしょうか？ ここで注意が必要なのは、Population OLS は実際には実行できず、事例数が限られたデータ上での OLS のみが可能なことです。

以下では 200 事例のデータについて、$f(X) = \beta_0 + \beta_1 X$ $(Y \sim X)$ および $f(X) = \beta_0 + \beta_1 X_1 + .. + \beta_8 X^8$ を OLS 推定しました。

[*1] 多重共線性の発生など、例外的なケースは存在します。

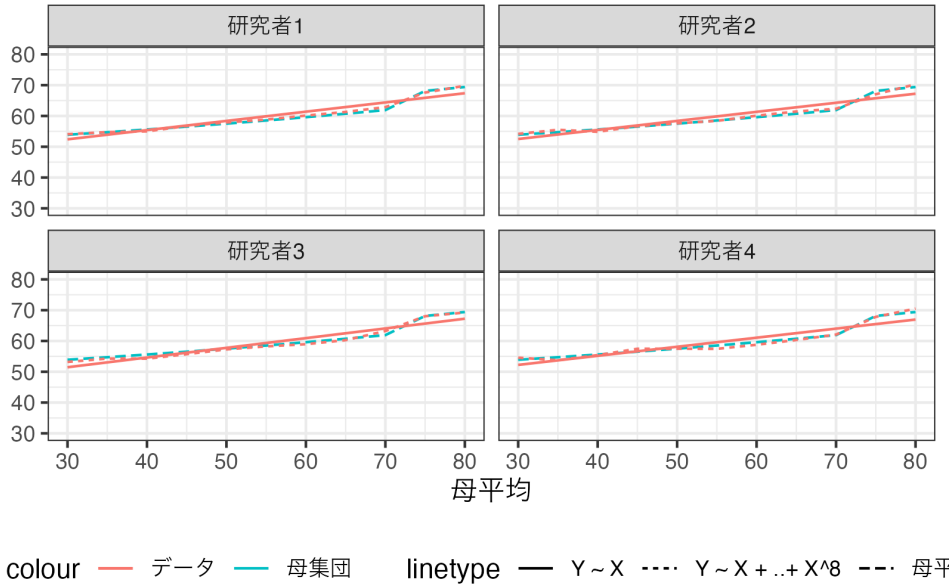


母平均は青の点線で示しています。データ上での OLS の結果は、複雑なモデルについては赤の点線、シンプルなモデルについては赤の実線で示しています。

複雑なモデルを推定した結果、単純なモデルよりも、母平均から大きく乖離した箇所が散見されます。さらに複雑なモデルについての推定結果は、研究者間で大きなばらつきが見られます。

これはモデルを複雑化した結果、データ上の平均値に近づくことに原因があります (Section 3.3)。限られた事例数のもとでは、小規模な事例のみから計算された平均値が発生します。このような小規模集計を避けるために、線型モデルを活用した集計を行います。しかしながらモデルを複雑化すると、この集計が再び上手くいかなくなります。複雑なモデルを OLS で推定すると、小規模な事例の偶然の偏りを強く反映してしまい、結果母平均から大きく乖離してしまうリスクが高くなります。

この問題は、大規模事例が活用できる場合は、発生しにくくなります。例えば 20000 事例について、OLS 推定を行った結果は以下です。



母平均は青の点線で示しています。データ上での OLS の結果は、複雑なモデルについては赤の点線、シンプルなモデルについては赤の実線で示しています。

複雑なモデルの推定結果は、母平均をよりよく近似していることが確認できます。対して単純なモデルは、依然として母平均から乖離しています。

このようなモデルの複雑さをめぐる問題は、Bias-Variance トレードオフと呼ばれてきました。大規模事例を用いることができるのであれば、複雑なモデルが母平均を上手く近似できます。事例数が少ない場合、複雑なモデルは母平均から大きく乖離してしまう可能性が高く、「妥協的に」単純なモデルを推定することが現実的です。

応用上の問題は、適切なモデルの複雑さについて、理論的な示唆が限られていることです。次の章では、複雑なモデルであったとしても適切に推定できる手法である、LASSO を紹介します。

4.4　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
  - estimatr: OLS 推定 + 頑健な信頼区間の計算
  - dotwhisker: 信頼区間の可視化

データを取得します。

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

lm_robust 関数を用いて OLS を推定します。

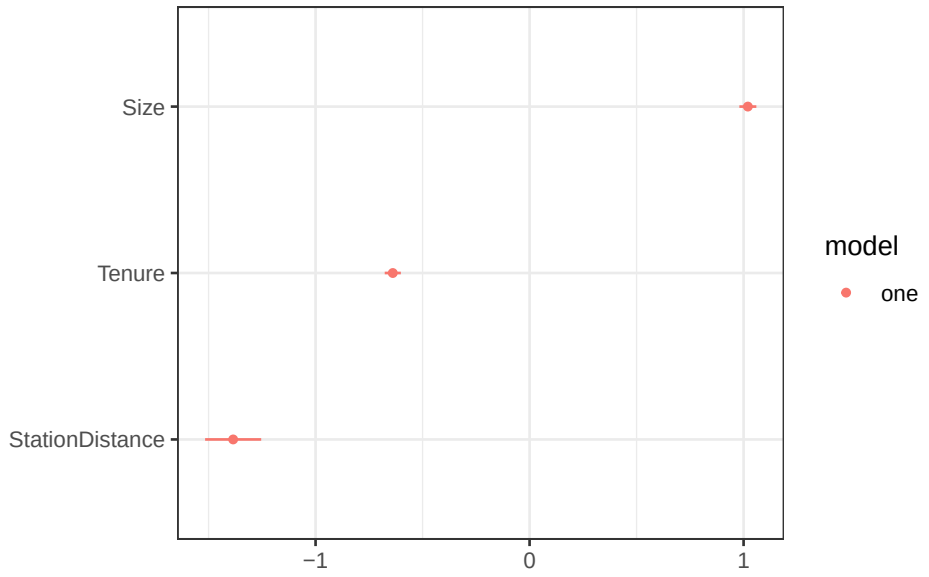
```
OLS = estimatr::lm_robust(
  Price ~ Size + Tenure + StationDistance,
  Data,
  alpha = 0.05 # 95% 信頼区間を計算
  ) # OLS

OLS
```

```
                  Estimate Std. Error   t value      Pr(>|t|)   CI Lower
(Intercept)     19.7205516 0.62170089  31.72032 3.664958e-205 18.5018088
Size             1.0199061 0.02020485  50.47828  0.000000e+00  0.9802978
Tenure          -0.6391756 0.01919741 -33.29488 3.014245e-224 -0.6768090
StationDistance -1.3850527 0.06684087 -20.72165  2.316510e-92 -1.5160833
                  CI Upper   DF
(Intercept)     20.9392944 6374
Size             1.0595144 6374
Tenure          -0.6015422 6374
StationDistance -1.2540221 6374
```

Estimate が β の推定値を、CI Lower/CI Upper が信頼区間の下限/上限を示します。例えば、推定された線型モデルにおいて、Size(部屋の広さ) と平均取引価格は正の関係性があります。さらに信頼区間は $[18.5, 20.9]$ なので、Population OLS においても正の関係性があることを強く示唆しています。

上記の推定結果は、dotwhisker 内の dwplot 関数を用いて可視化できます。



各点が推定値、横棒が信頼区間を図示しています。

第 5 章

LASSO

複雑なモデルを適切に推定する方法として、LASSO (Tibshirani 1996) を紹介します。LASSO は、罰則付き回帰と呼ばれる枠組みの一つの手法です[*1]。OLS と同様に線型予測モデルを推定しますが、データへの当てはまりだけでなく、モデルの複雑性も抑制することも目指します。

5.1 推定方法

例として、以下のモデルの推定を目指します。

$$\beta_1 Size \times \text{板橋区ダミー} + .. + \beta_6 Size^6 \times \text{板橋区ダミー}$$
$$+.. + \beta_{132} Size \times \text{中央区ダミー} + .. + \beta_{138} Size^6 \times \text{中央区ダミー}$$

合計 138 個のパラメタがあり、事例数次第では、OLS による推定は困難です。これは OLS 推定が、以下を最小化するように β を推定しているためです。

$$(Y - \text{モデル})^2 \text{のデータ上の平均}$$

β の数が多いと、データへの不適合度 $(Y - \text{モデル})^2$ をいくらでも低下させられるため、複雑なモデルを推定すると**データへの**過剰な適合 (母平均からの乖離) を引き起こします。

LASSO 推定では、β の値を以下を最小化するように決定します。

定義 LASSO

$$(Y - \text{予測値})^2 \text{のデータ上の平均}$$

[*1] 他の手法として、Ridge や Best subset selection などがあります。詳細は、James et al. (2021) などを参照ください。

$$+ \underbrace{\lambda}_{Tunning\ Parameter} \times(\beta_1\text{の絶対値} + ..)$$

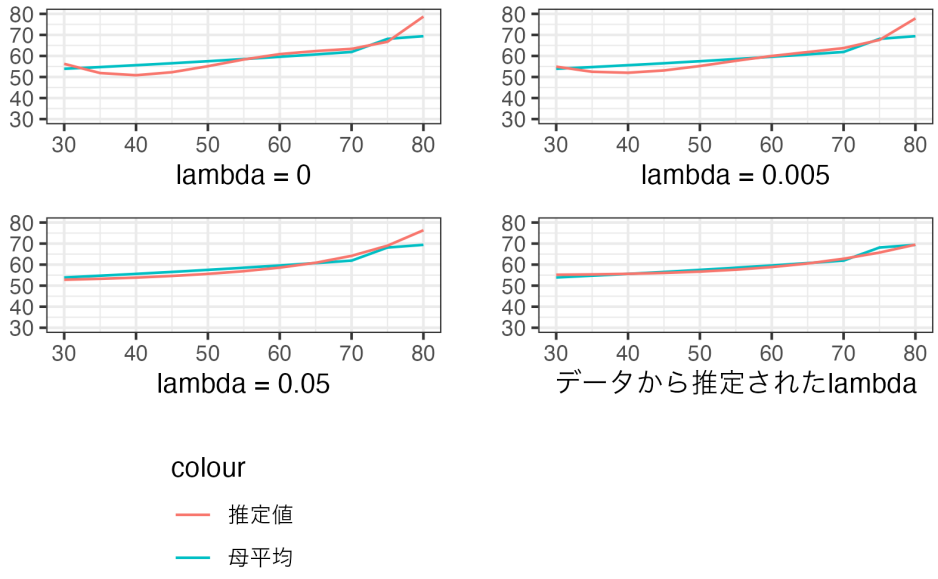
λ は、データへの当てはまりではなく、モデルの予測性能を高めるように決定する。具体的には、交差検証を用いる方法 (Tibshirani 1996)、情報基準などの理論的な評価指標を用いる方法などがある (Belloni, Chernozhukov, and Hansen 2014; Taddy 2017) 。

λ に応じて、予測モデルがどのように変化するのか考えてみます。λ を変化させることで、予測モデルは、単純平均と複雑なモデルの OLS の間で変化することになります。$\lambda = 0$ であれば、OLS と全く同じモデルを推定します。よって複雑な線型モデルを推定した場合は、データ上の平均値に近いモデルとなります。λ を非常に大きい値を設定した場合、$\beta_1 = \beta_2 = .. = 0$ となります。この場合は β_0 をデータに当てはまるように推定することになります。

以下の数値例は、200 事例からなるデータについて、LASSO により以下のモデルを推定しました

$$\beta_0 + \beta_1 Size + .. + \beta_6 Size^6$$

λ については、0,0.005,0.05、および赤池情報基準により選ばれた値 (Taddy 2017) を使用した結果を図示しています。



母平均を青線、LASSO による推定結果を赤線で示しています。

$\lambda = 0$ に比べると、λ の値が大きくなるにつれ、モデルが単純な曲線に近づいていることが確認できます。また Taddy (2017) に基づいて設定された λ のもとでは、かなり単純化

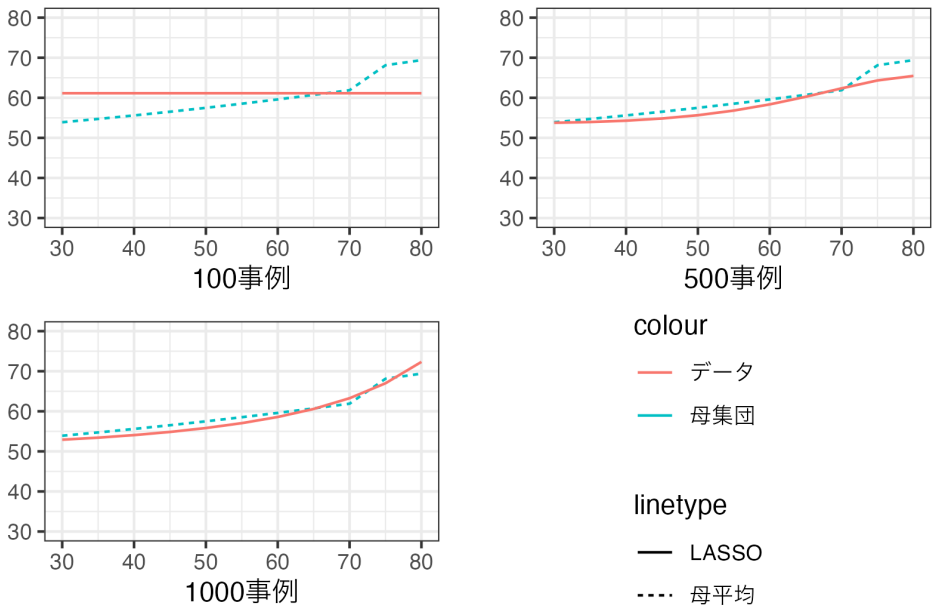
されたモデルが推定されたことも確認できます。

> **練習問題**
>
> λ は、β と異なり、データへの当てはまりを最大化するように決定できません。なぜでしょうか？

5.1.1 事例数の拡大

推定結果は、一般に事例数に強く影響を受けます。特に LASSO などの機械学習の方法においては、データの特徴により強く依存します。

以下の数値例では、事例数を 100 事例から 1000 事例まで増やし、$Y \sim Size + .. + Size^{10}$ を LASSO で推定しています (λ は Taddy (2017) の方法で設定しています)。



事例数の増加とともに、モデルの複雑性が、「自動調整」されていることが確認できます。100 事例では、単純平均値が推定されており、極めて単純なモデルが採用されています。事例が増えると、モデルの傾きに加えて、「曲がり方」も変化しており、モデルが複雑化しています。

5.2 信頼区間

LASSO によって推定されたパラメタについて、信頼区間を計算する方法は盛んに議論されているものの、筆者の知る限り、現状確立された方法は存在しません[*2]。

これは LASSO 以外の「機械学習」の手法についても同様です。一般にデータと推定値との関係性は、伝統的な推定方法と比べて、機械学習の方が複雑になります。このため、理論的な関係性を導くのが難しく、信頼区間の計算方法の確率が困難となっています。

5.3 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
  - gamlr: LASSO

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

X = model.matrix(
  ~ 0  +
    (Size + Tenure + StationDistance + District)**2 + # 交差項
    I(Size^2) + I(Tenure^2) + I(StationDistance^2), # 二乗項
  Data
) # X の作成

X = scale(X) # 標準化

colnames(X) # X に格納されている変数の確認
```

```
 [1] "Size"                          "Tenure"
 [3] "StationDistance"               "District 世田谷区"
 [5] "District 中央区"               "District 中野区"
 [7] "District 北区"                 "District 千代田区"
 [9] "District 台東区"               "District 品川区"
[11] "District 大田区"               "District 文京区"
[13] "District 新宿区"               "District 杉並区"
```

[*2] 具体的な議論については、Chernozhukov, Hansen, and Spindler (2015); Kuchibhotla, Kolassa, and Kuffner (2022) などを参照

```
 [15] "District 板橋区"                    "District 江戸川区"
 [17] "District 江東区"                    "District 渋谷区"
 [19] "District 港区"                      "District 目黒区"
 [21] "District 練馬区"                    "District 荒川区"
 [23] "District 葛飾区"                    "District 豊島区"
 [25] "District 足立区"                    "District 墨田区"
 [27] "I(Size^2)"                          "I(Tenure^2)"
 [29] "I(StationDistance^2)"               "Size:Tenure"
 [31] "Size:StationDistance"               "Size:District 中央区"
 [33] "Size:District 中野区"               "Size:District 北区"
 [35] "Size:District 千代田区"             "Size:District 台東区"
 [37] "Size:District 品川区"               "Size:District 大田区"
 [39] "Size:District 文京区"               "Size:District 新宿区"
 [41] "Size:District 杉並区"               "Size:District 板橋区"
 [43] "Size:District 江戸川区"             "Size:District 江東区"
 [45] "Size:District 渋谷区"               "Size:District 港区"
 [47] "Size:District 目黒区"               "Size:District 練馬区"
 [49] "Size:District 荒川区"               "Size:District 葛飾区"
 [51] "Size:District 豊島区"               "Size:District 足立区"
 [53] "Size:District 墨田区"               "Tenure:StationDistance"
 [55] "Tenure:District 中央区"             "Tenure:District 中野区"
 [57] "Tenure:District 北区"               "Tenure:District 千代田区"
 [59] "Tenure:District 台東区"             "Tenure:District 品川区"
 [61] "Tenure:District 大田区"             "Tenure:District 文京区"
 [63] "Tenure:District 新宿区"             "Tenure:District 杉並区"
 [65] "Tenure:District 板橋区"             "Tenure:District 江戸川区"
 [67] "Tenure:District 江東区"             "Tenure:District 渋谷区"
 [69] "Tenure:District 港区"               "Tenure:District 目黒区"
 [71] "Tenure:District 練馬区"             "Tenure:District 荒川区"
 [73] "Tenure:District 葛飾区"             "Tenure:District 豊島区"
 [75] "Tenure:District 足立区"             "Tenure:District 墨田区"
 [77] "StationDistance:District 中央区"    "StationDistance:District 中
野区"
 [79] "StationDistance:District 北区"      "StationDistance:District 千代
田区"
 [81] "StationDistance:District 台東区"    "StationDistance:District 品
川区"
 [83] "StationDistance:District 大田区"    "StationDistance:District 文
京区"
```

```
[85] "StationDistance:District 新宿区"   "StationDistance:District 杉
並区"
[87] "StationDistance:District 板橋区"   "StationDistance:District 江戸川
区"
[89] "StationDistance:District 江東区"   "StationDistance:District 渋
谷区"
[91] "StationDistance:District 港区"     "StationDistance:District 目
黒区"
[93] "StationDistance:District 練馬区"   "StationDistance:District 荒
川区"
[95] "StationDistance:District 葛飾区"   "StationDistance:District 豊
島区"
[97] "StationDistance:District 足立区"   "StationDistance:District 墨
田区"
```

合計 101 個のパラメタ推定を目指します。

gamlr パッケージ内の gamlr 関数を用いて LASSO 推定をします。

```
LASSO = gamlr::gamlr(
  y = Data$Price, # Y の指定
  x = X
) # LASSO 推定
```

推定された値は、以下で示します。“.” は、厳密に 0 であることを意味しています。

```
coef(LASSO) # 推定値
```

```
99 x 1 sparse Matrix of class "dgCMatrix"
                                        seg100
intercept                         42.705221072
Size                              22.085283332
Tenure                            -0.949260232
StationDistance                    .
District 世田谷区                  0.885399681
District 中央区                    .
District 中野区                    .
District 北区                      .
District 千代田区                  .
District 台東区                    .
District 品川区                    .
```

District 大田区	.
District 文京区	.
District 新宿区	.
District 杉並区	.
District 板橋区	.
District 江戸川区	.
District 江東区	.
District 渋谷区	.
District 港区	.
District 目黒区	0.145531303
District 練馬区	.
District 荒川区	.
District 葛飾区	.
District 豊島区	.
District 足立区	.
District 墨田区	.
I(Size^2)	5.098523148
I(Tenure^2)	0.764221115
I(StationDistance^2)	.
Size:Tenure	-9.504525202
Size:StationDistance	-3.840468008
Size:District 中央区	2.207591886
Size:District 中野区	.
Size:District 北区	-0.947403269
Size:District 千代田区	4.536157097
Size:District 台東区	.
Size:District 品川区	1.556391250
Size:District 大田区	-0.999011985
Size:District 文京区	1.658065622
Size:District 新宿区	2.554318454
Size:District 杉並区	0.092031271
Size:District 板橋区	-2.322997154
Size:District 江戸川区	-2.265860521
Size:District 江東区	-0.573435169
Size:District 渋谷区	5.671810968
Size:District 港区	12.715565955
Size:District 目黒区	2.106822926
Size:District 練馬区	-1.492913960
Size:District 荒川区	-1.064818067

```
Size:District 葛飾区                -2.532339361
Size:District 豊島区                 0.818647790
Size:District 足立区                -3.474390892
Size:District 墨田区                -0.677389749
Tenure:StationDistance               0.285838918
Tenure:District 中央区                .
Tenure:District 中野区                .
Tenure:District 北区                  .
Tenure:District 千代田区            -1.110147900
Tenure:District 台東区                0.400329988
Tenure:District 品川区               -0.002439637
Tenure:District 大田区                .
Tenure:District 文京区                .
Tenure:District 新宿区               -0.219994692
Tenure:District 杉並区                .
Tenure:District 板橋区                .
Tenure:District 江戸川区              .
Tenure:District 江東区                .
Tenure:District 渋谷区               -1.158716799
Tenure:District 港区                 -2.258199491
Tenure:District 目黒区                .
Tenure:District 練馬区                .
Tenure:District 荒川区                .
Tenure:District 葛飾区                .
Tenure:District 豊島区                .
Tenure:District 足立区                .
Tenure:District 墨田区                .
StationDistance:District 中央区       .
StationDistance:District 中野区       0.006454483
StationDistance:District 北区         .
StationDistance:District 千代田区  -0.259898503
StationDistance:District 台東区       .
StationDistance:District 品川区       .
StationDistance:District 大田区       .
StationDistance:District 文京区       .
StationDistance:District 新宿区       0.099488399
StationDistance:District 杉並区       .
StationDistance:District 板橋区       .
StationDistance:District 江戸川区     .
```

```
StationDistance:District 江東区     .
StationDistance:District 渋谷区     .
StationDistance:District 港区      -2.157523041
StationDistance:District 目黒区     0.217234860
StationDistance:District 練馬区     .
StationDistance:District 荒川区     .
StationDistance:District 葛飾区     .
StationDistance:District 豊島区     .
StationDistance:District 足立区     .
StationDistance:District 墨田区     .
```

第 6 章

予測分析への応用

母平均のモデルの有力な応用例として、Yの予測問題への活用があります。予測を目的とする研究では、観察できる変数 X から意思決定に必要だが欠損している情報 Y を予測するモデル（予測モデル）の推定を目指します。

実例は、以下があげられます。

- 視聴履歴や” いいね数” (X) $\underbrace{\rightarrow}_{\text{予測モデル}}$ 好む未視聴動画 (Y)
- メールの文名や件名 (X) $\underbrace{\rightarrow}_{\text{予測モデル}}$ 迷惑メール (Y)
- 事業の内容や財務状況 (X) $\underbrace{\rightarrow}_{\text{予測モデル}}$ デフォルトリスク (Y)

統計学や機械学習においては、もし Y, X が共に観察できるデータを活用できるのであれば、予測モデルの推定方法がある程度確立されています。

6.1 推定目標

本章では、データと同じ母集団から**新たに**抽出された事例を、予測するモデルの推定を目指します。予測性能は、新たな事例が抽出される母集団における平均二乗誤差で測定します

$$\text{母集団における平均二乗誤差} = (Y - \text{予測値})^2\text{の母集団における平均}$$

母集団における平均二乗誤差は、直接計算することは不可能です。ただし Section 6.2 で紹介する通り、推定することは可能です。

本説では、実際に推定する前に、予測モデルが持つ基本的な性質を確認します。

6.1.1　完璧な予測は不可能

予測研究における究極的な目標の一つは、Y を完璧に予測するモデルの推定です。しかしながら多くの応用で、この目標には到達することができません。予測モデルは、ある X の組み合わせについて、一つの予測値のみを出力します。このため**母集団**において、同じ X 内で Y の値にばらつきがあれば、予測が外れる事例は必ず存在します。

完璧な予測には、Y の全ての決定要因を X として観察し、X 内での個人差をなくす必要があります。しかしながら人間行動や社会的な事象などの社会的変数の決定要因は無数に存在し、その多くは観察が難しいと考えられます。結果、社会的変数について、完璧な予測は不可能と考えられます。

6.1.2　理想の予測モデル

最も高い予測性能 (母集団における平均二乗誤差が最小化) を達成するモデルは、母平均であることが証明できます。このため予測研究においては、データから推定した予測モデルを理想の予想モデルである母平均に極力近づけることが、実質的な目標となります。具体的には以下を目指します

$$\underbrace{X\text{内での}Y\text{の母平均}}_{\text{理想の予測モデル}} \simeq \text{予測モデル}$$

OLS や LASSO は、母平均のモデルを指定しており、理想の予測モデル推定を目指すための有力な方法であると考えられます。

6.2　予測性能の測定

予測を目指す分析では、推定された予測モデルの性能を評価することが重要となります。OLS や LASSO により推定されたモデルの性質について、多くの理論研究が蓄積されています。しかしながら現状、幅広いデータや状況において、一貫して高い予測性能を生み出す方法は存在しません。また実際の予測力を理論のみに基づいて、算出することは困難です。このため複数の予測モデルを” 試作” し、データを用いてその性能を比較することが求められます。

最もシンプルな評価方法は、サンプル分割です。

❗ サンプル分割法の定義

1. データをランダムに訓練データとテストデータに分割する
- 訓練/テスト間での事例数の比率については、8 対 2 や 95 対 5 が (経験則と

OLS	OLS| 交差項 + 高次項	LASSO
369	316	303

して) 推奨されることが多い。
2. 訓練データのみでモデルを試作する
3. テストデータへの当てはまりを (平均二乗誤差などで) 評価する

実際の取引データに適用した結果は以下です。6378 事例のうち、ランダムに選んだ半分を訓練、残り半分をテストに用いました。テストした推定方法は以下です。

- **OLS:** 取引価格 ~ 部屋の広さ + 立地 + 取引年
- **OLS (含む交差項 + 高次項):** 取引価格 ~ 部屋の広さ + 立地 + 取引年 + 交差項 + 高次項 (6 次まで)
- **LASSO (含む交差項 + 高次項):** 取引価格 ~ 部屋の広さ + 立地 + 取引年 + 交差項 + 高次項 (6 次まで)

推定された予測モデルの評価結果は以下となりました。

LASSO が最も予測性能が高く、単純な OLS と比較し、0.179 パーセントほど平均二乗誤差を削減しています。対して、交差項と高次項を追加した OLS と LASSO を比較した場合、0.041 パーセントほどの改善にとどまります。これはモデルの複雑さに比べて、事例数が多く、LASSO によるモデル単純化の恩恵が限定的であることを示しています。

6.3 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
  - gamlr: LASSO

6.3.1 準備

データとその事例数を取得し、データをランダムに分割します。

```
set.seed(111)

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

```
X = model.matrix(
  ~ 0  +
    (Size + Tenure + StationDistance + District)**2 + # 交差項
    I(Size^2) + I(Tenure^2) + I(StationDistance^2), # 二乗項
  Data
) # X の作成

X = scale(X) # 標準化

N = nrow(Data) # 事例数の取得

Group = sample(
  1:2,
  N, # 事例数
  replace = TRUE, # 復元抽出を指定
  prob = c(0.8,0.2) # "1"が 8 割、"2"が 2 割
) # サンプル分割のために 1 または 2 をランダムに発生
```

lm 関数を用いて OLS, gamlr 関数を用いて LASSO 推定をします。また LASSO 推定は、複雑なモデルを推定に利点を持つため、交差項と二乗項までを導入したモデルも推定しています。

```
LASSO = gamlr::gamlr(
  y = Data$Price[Group == 1],
  x = X[Group == 1,]) # LASSO

OLS = lm(
  Price ~ Size + Tenure + StationDistance + District,
  Data,
  subset = Group == 1
)

OLS_Long = lm(
  Price ~ +
    (Size + Tenure + StationDistance + District)**2 +
    I(Size^2) + I(Tenure^2) + I(StationDistance^2),
  Data,
  subset = Group == 1
)
```

```
HatLASSO = predict(LASSO,X)
HatOLS = predict(OLS,Data)
HatOLS_Long = predict(OLS_Long,Data)

mean((Data$Price - HatLASSO)[Group == 2]^2)
```

```
[1] 154.7488
```

```
mean((Data$Price - HatOLS)[Group == 2]^2)
```

```
[1] 265.1663
```

```
mean((Data$Price - HatOLS_Long)[Group == 2]^2)
```

```
[1] 149.513
```

6.4 まとめ

典型的な予測問題においては、データを用いた予測モデルの推定と評価、という二つの作業が求められます。一般にこの作業を同じデータで行うと多くの問題が発生します。本章では、データをランダムに 2 分割し、片方のデータで予測を、残りのデータで評価を行う方法を紹介しました。

評価値としては平均二乗誤差を用いました。現在、他の評価方法も盛んに研究されています (Angelopoulos, Bates, et al. (2023) などを参照してください)。

第 7 章

回帰木モデル

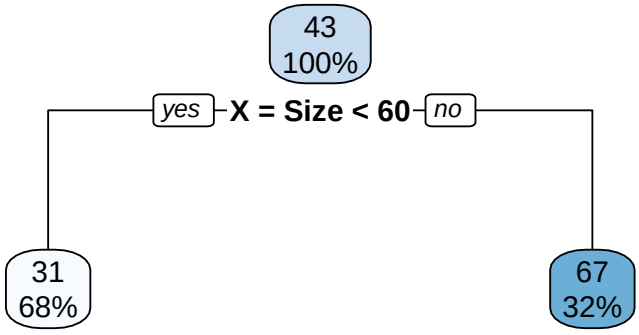
本章では線型モデルの有力な代替案である、回帰木モデルを紹介します。標準的な回帰木は、” サブグループ平均値” を母平均の推定値とします。

回帰木モデルにおける最大の論点は、サブグループをどのように定義するのか、にあります。伝統的には、研究者が背景知識などを用いて定式化してきました。近年では、サブグループの定義もデータ主導で行う方法が注目されています。特に**モデル集計**と呼ばれる手法も用いることで、母平均の近似精度を大きく改善することが期待されます。

7.1　伝統的な方法

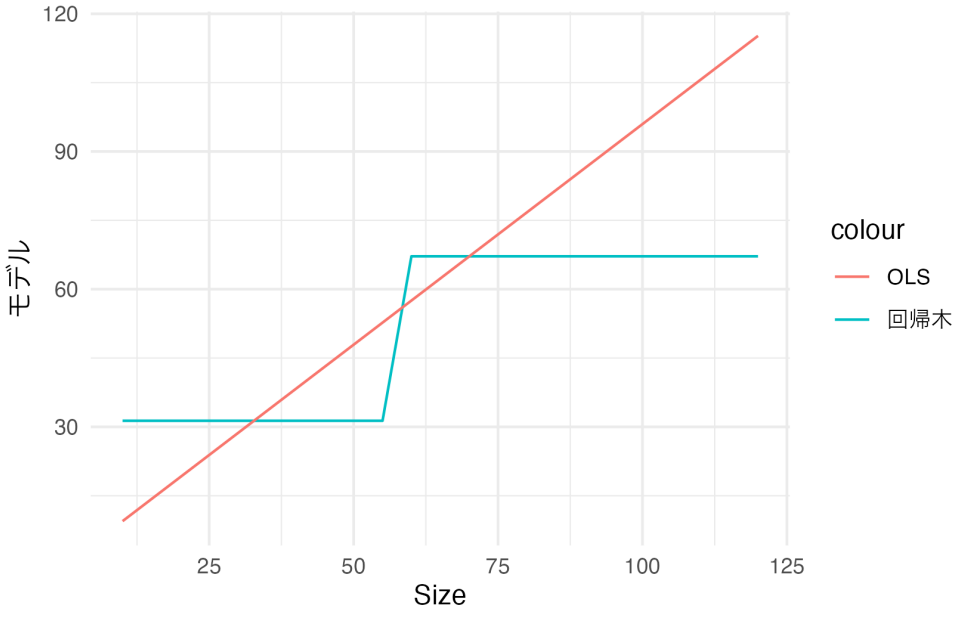
例えば、$X =$ [部屋の広さ]、$Y =$ 取引価格、について、母平均 $E[Y \mid X]$ を推定する際に、サブグループ $\{$ 部屋の広さが 60 以上, 部屋の広さが 60 以下 $\}$ に分け、各サブグループごとに平均取引価格を計算し推定とすることができます。

このようなモデルは、以下のように樹形図として表現できます。



一番上のボックスには、データ全体の平均取引価格 (43) とデータ全体に占める事例数の割合 (100 %) を示しています。下のボックスは、各サブグループの平均取引価格と事例数の割合を示しています。部屋の広さが 60 以下であれば”yes”、以上であれば”no” です。例えば、60 以下のサブグループにおけう平均取引価格は 31、事例数の割合は 68 % となります。

また散布図に示すと以下のように、“階段状” のモデルとなります。赤線が回帰モデル、青線は比較対象として OLS ($Price \sim \beta_0 + \beta_1 \times Size$) を示しています。



7.2　データ主導の方法

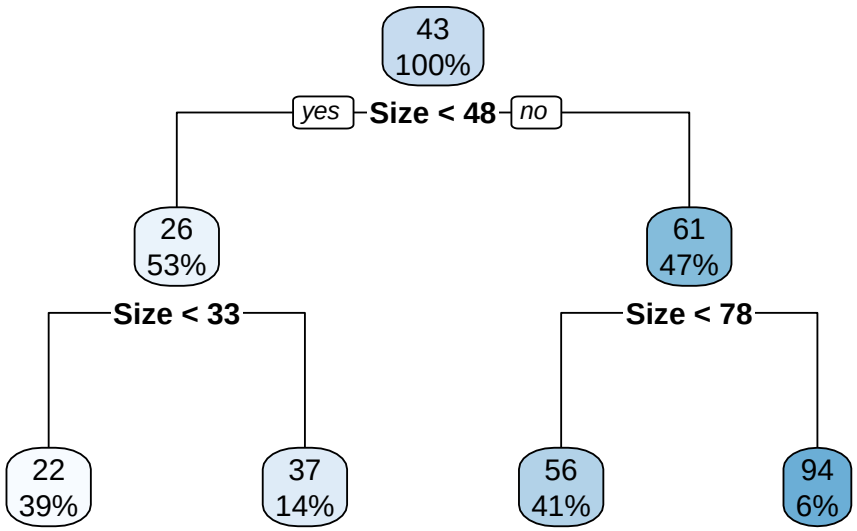
伝統的な方法の応用上の問題は、サブグループの定義を研究者が行う必要があることです。研究課題によっては、このような研究者主導のサブグループ分けは困難です。

この問題について、データ主導のサブグループ分けが提案されています。最も代表的な方法は、データへの当てはまりの良さを最大にするようにサブグループを定義する方法です。この場合は、最大分割回数や各サブグループの事例数の下限値を設定した後、二乗誤差の平均値を最小化するようにサブグループを定義する方法が一般的です。

以下では rpart 関数を用いて、最大 2 分割、最小事例数を 50 として、回帰木を推定しました。Y は取引価格、$X = [Size]$ です。

```
Model = rpart::rpart(
    Price ~ Size,
    data = Data,
    control = rpart::rpart.control(
      maxdepth = 2,
      minbucket = 50,
      minsplit = 1
      )
    ) # 回帰木の推定

rpart.plot::rpart.plot(Model) # 可視化
```



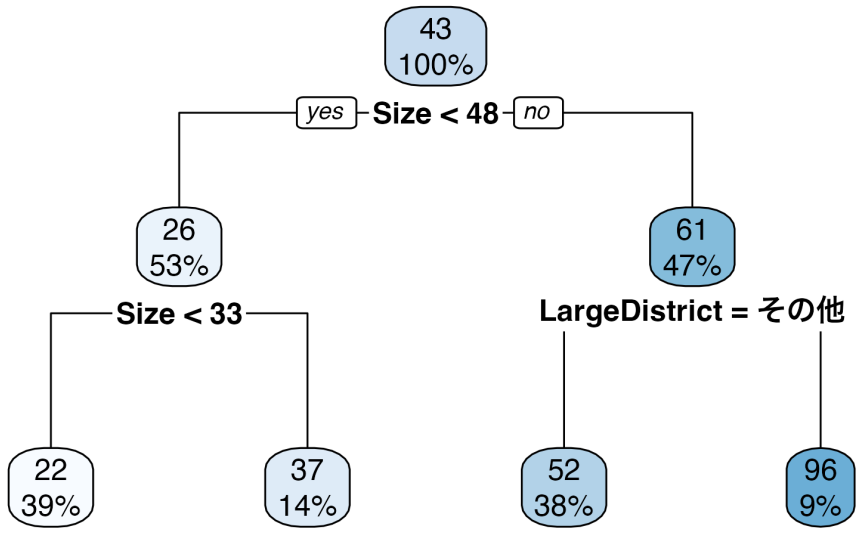
rpart 関数は、「貪欲なアルゴリズム」を用いて、2 グループへの分割を繰り返します。一番最初の分割では、Size が 48 以上か否かでサブグループが定義されました。これは 48 以上か否かで分割しサブグループ平均を計算するモデルが、最もデータへの当てはまりが良いためです。

48 以下の物件については、33 以上か否かで 2 回目の分割が行われました。この理由は、1 回目の分割と同様に、33 以上か否かで分割したモデルのデータへの当てはまりが良いためです。同様に 48 以上については、78 以上か否かで分割されます。

回帰木は、複数の X を用いたモデルも生成できます。以下では、Size に加え、築年数 (Tenure) および物件の立地 (LargeDistrict $=$ [中心 6 区、その他]) も加えたモデルを推定しています。すなわち $X = [Size, Tenure, LargeDistrict]$ となります。

```
Model = rpart::rpart(
    Price ~ Size + Tenure + LargeDistrict,
    data = Data,
    control = rpart::rpart.control(
      maxdepth = 2,
      minbucket = 50,
      minsplit = 1
      )
    ) # 回帰木の推定

rpart.plot::rpart.plot(Model) # 可視化
```



結果、Size が 48 以上の物件については、立地が中心 6 区か否かで 2 回目の分割が行われました。

7.3 過剰適合への対処

回帰木における複雑性は、最大分割回数や各サブグループの事例数の下限値などに操作できます。最大分割回数を増やし、事例数の下限を減らせば、モデルはより複雑化します。複雑なモデルは、事例数が十分にあれば、母平均の特徴をよりよく捉えることができます。一方で OLS と同様に、複雑な回帰木モデルを推定すると、データへの適合度は高まる一方で、母平均からは乖離する傾向が生じます。

このような問題に対しては、LASSO 同様に、モデルを適切に単純化する方法が考えられます。代表的な方法としては、「剪定 (Pruning)」などが挙げられます。詳細は、James et al. (2021) などを参照ください。

次節では、母平均の近似を目指す場合により有力な方法である、モデル集計を紹介します。

7.4 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
  - rpart (R に同梱): 回帰木の推定

– rpart.plot：回帰木の可視化

データを取得します。

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

rpart 関数を用いて OLS を推定します。

```
Tree = rpart::rpart(
  Price ~ Size + Tenure + StationDistance + LargeDistrict, # Y ~ X
  Data, # 使用するデータの指定
  control = rpart::rpart.control(
    maxdepth = 3, # 最大分割回数 = 2
    minsplit = 50, # 50 以下になったら分割を停止
    minbucket = 50 # 50 以下のサブグループを作らない
  )
  )

Tree
```

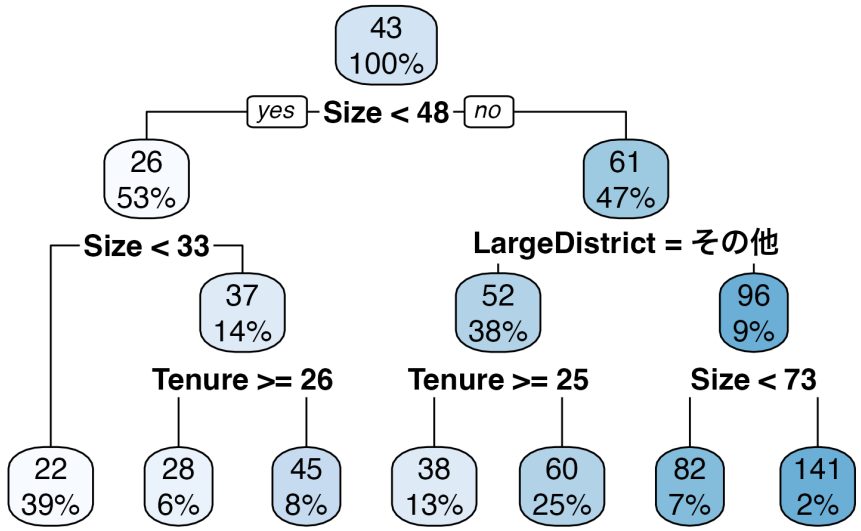
```
n= 6378

node), split, n, deviance, yval
      * denotes terminal node

 1) root 6378 5935190.00  42.70522
   2) Size< 47.5 3359  503187.50  26.41390
     4) Size< 32.5 2458  147489.90  22.38031 *
     5) Size>=32.5 901  206606.60  37.41787
      10) Tenure>=25.5 386  43399.57  27.53238 *
      11) Tenure< 25.5 515  97213.62  44.82718 *
   3) Size>=47.5 3019 3548597.00  60.83127
     6) LargeDistrict=その他 2430 1257966.00  52.28255
      12) Tenure>=24.5 837  218600.20  38.45950 *
      13) Tenure< 24.5 1593  795403.00  59.54551 *
     7) LargeDistrict=中心 6 区 589 1380391.00  96.10017
      14) Size< 72.5 446  412038.60  81.80269 *
      15) Size>=72.5 143  592832.50 140.69230 *
```

rpart.plot 関数を用いて可視化します。

```
rpart.plot::rpart.plot(Tree)
```



Reference

Angelopoulos, Anastasios N, Stephen Bates, et al. 2023. "Conformal Prediction: A Gentle Introduction." *Foundations and Trends® in Machine Learning* 16 (4): 494–591.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press.

Aronow, Peter M, and Benjamin T Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects After Selection Among High-Dimensional Controls." *Review of Economic Studies* 81 (2): 608–50.

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, Christian Hansen, and Martin Spindler. 2015. "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach." *Annu. Rev. Econ.* 7 (1): 649–88.

Ding, Peng. 2024. "Linear Model and Extensions." *arXiv Preprint arXiv:2401.00649*.

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

Kuchibhotla, Arun K, John E Kolassa, and Todd A Kuffner. 2022. "Post-Selection Inference." *Annual Review of Statistics and Its Application* 9 (1): 505–27.

Lin, Hanti. 2024. "To Be a Frequentist or Bayesian? Five Positions in a Spectrum." *Harvard Data Science Review* 6 (3).

Taddy, Matt. 2017. "One-Step Estimator Paths for Concave Regularization." *Journal of Computational and Graphical Statistics* 26 (3): 525–36.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58 (1): 267–88.