データ分析: 入門

川田恵介

2026-01-21

Table of contents

第 1 章

はじめに

本ページでは、データ分析における基本的な手法の一つである「母平均」および「Population OLS」の推定方法について、その基本概念とともに紹介します。

特に、「推定対象として定義された母集団上での計算と同じ操作を、実際のデータ上で行う」というシンプルかつ応用の範囲が広い推定の考え方に注目します。このアプローチは、プラグイン原理（plug-in principle）やアナログ原理（analogy principle）として知られており、より一般的な推定方法の一例としても位置づけられます。

また、OLS 推定においては、モデルの定式化が結果の解釈や推定精度に大きく影響することにも注意が必要です。現代の実証研究では、研究目標に応じて適切なモデルを構築することが極めて重要であり、本ページではその点についても強調していきます。

第 2 章

データの要約

データ分析において、結果を人間が理解できる形で要約することは不可欠です。これは、分析結果は予測や推薦、社会構造に関する示唆など多様な形で提示されますが、最終的には意思決定者が理解できる形式で示す必要があるためです。

> ❗ 到達目標
>
> - データの要約が必要な理由の理解
> - 「条件付き分布 → 条件付き平均値 → 線型モデル」の順番で、変数間の関係性が要約できることの理解

2.1　分布による要約

具体的例を示すために、まず AER package から CPSSW04 データを読み込み、一部の事例を表示します。

Table2.1

```
library(tidyverse) # tidyverse パッケージの読み込み

data("CPSSW04", package = "AER") # データの読み込み

head(CPSSW04) # 一部の事例の抜き出し
```

```
  earnings     degree gender age
1 34.61538   bachelor   male  30
2 19.23077   bachelor female  30
3 13.73626 highschool female  30
4 19.23077   bachelor female  30
5 19.23077   bachelor   male  25
6 38.46154   bachelor female  32
```

CPSSW04 はアメリカの労働者についての調査であり、四つの変数 (所得: earnings, 学位: degree, 性別: gender, 年齢: age) が観察できます。また合計で 7986 事例が含まれています。

このような巨大なデータを直接理解できる人間は、おそらく存在しないでしょう。

そこで、最も基本的な要約方法は分布表を作成してみます。分布とは、変数の値が同一である事例数を示します。例えば、count 関数を用いると、学位と性別の分布は次のように計算できます。

```
count(
  CPSSW04,
  degree,
  gender)
```

```
     degree gender    n
1 highschool   male 2772
```

```
2 highschool female 1574
3   bachelor   male 1901
4   bachelor female 1739
```

n に事例数が表示されています。例えば、男性で高校卒の事例は 2,772 件存在することがわかります。

割合を加えることで、データ全体に占める比率も把握できます。

```
      degree gender    n      割合
1 highschool   male 2772 0.3471074
2 highschool female 1574 0.1970949
3   bachelor   male 1901 0.2380416
4   bachelor female 1739 0.2177561
```

CPSSW04 において、全体の 1/3 程度が高校卒の男性であることなどがわかります。

2.1.1 分布の限界

CPSSW04 には、earnings や age などの連続変数も含まれています。これらは CPSSW04 データの特徴を理解する上で、非常に重要な変数でしょう。しかしながら複数の連続変数を含めると、分布表は巨大になり、理解が困難です。

例えば、earnings, age, degree, gender について、度数分布表を作成すると、4,649 行の表が作成されます。このような巨大な分布表の理解は、非常に困難です。

なお少数の連続変数の分布に注目するのであれば、ヒストグラムなどの分布を可視化するツール (Data visualization) を利用することで、人間が認識しやすくできます。しかしながら複数の変数の分布を可視化することは、困難です。

2.2　平均値による要約

分布全体の特徴ではなく、個別の変数の特徴を把握する方法も数多く提案されています。中でも各変数の分布を、その特徴を表す統計量 (statistics) に要約する方法は有力です。

最も代表的なものは、平均値です。例えば earnings の平均値 (平均所得) は、以下の方法で計算結果として定義できます。

$$\text{平均所得} = \frac{\text{第1事例の所得} + \text{第2事例の所得} + ...}{\text{事例数}}$$

平均値は、**分布のみ**を用いて計算することも可能です。

$$\begin{aligned}\text{平均所得} = 1 \times &\text{「所得が1」の割合} \\ +2 \times &\text{「所得が2」の割合}\end{aligned} \tag{2.1}$$

無論、どちらの方法でも同じ値が計算されます。

CPSSW04 に含まれるすべての変数について、summary 関数は、他の統計量とともに、平均値を計算します。

Mean に表示される数値が平均値です。例えば、CPSSW04 データにおける平均年齢は 29.75 才です。

2.2.1　平均値の限界

平均値には、変数間の関係性を示さないという問題があります。例えば、Table 2.2 を見ても、earnings と age,degree,gender の間の関係性について、何もわからないでしょう。

変数間の関係性を把握する方法には、数多くの提案があります。ここでは条件付き母平均、および線型モデルを用いる方法を紹介します。

Table2.2

```
summary(CPSSW04)
```

```
    earnings             degree        gender          age       
 Min.   : 2.098   highschool:4346   male  :4673   Min.   :25.00  
 1st Qu.:10.684   bachelor  :3640   female:3313   1st Qu.:27.00  
 Median :14.904                                  Median :30.00  
 Mean   :16.771                                  Mean   :29.75  
 3rd Qu.:20.292                                  3rd Qu.:32.00  
 Max.   :61.058                                  Max.   :34.00  
```

2.3 条件付き平均値による要約

もし少数の目的変数と多数の説明変数の関係性を把握することが目標であれば、**条件付き分布**から計算できる**条件付き平均値**の利用が有効です。

条件付き分布は、ある特定の条件を満たす事例内での分布です。例えば、「大学卒内での earnings $= 1$ の割合」は以下で計算できます。

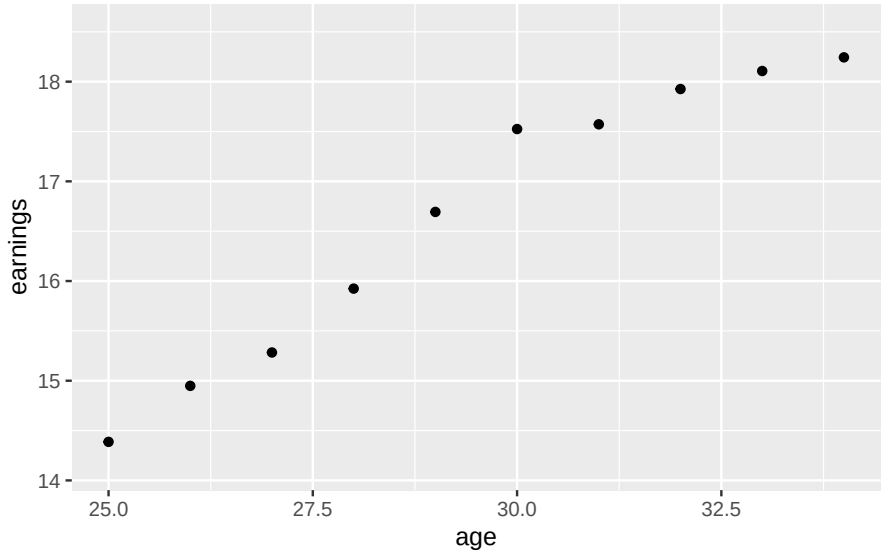
$$\text{大学卒内での「所得が1」の事例割合} = \frac{\text{大学卒かつ「所得が1」の事例割合}}{\text{大学卒の事例割合}}$$

条件付き平均値は、ある特定の条件を満たす事例内での平均値です。例えば、大学卒という条件を満たす平均年齢は以下のように計算できます。

$$\begin{aligned}&\text{大学卒の平均所得}\\ &= 1 \times \text{大学卒内での「所得が1」の割合}\\ &+2 \times \text{大学卒内での「所得が2」の割合} + \ldots\end{aligned}$$

例えば、特定の年齢ごとに条件付き平均賃金を示してみます。

```
ggplot(
  CPSSW04,
  aes(
    x = age,
    y = earnings
  )
) +
  stat_summary(
    geom = "point"
  )
```



横軸は年齢、縦軸は平均所得、各点が年齢ごとの条件付き平均賃金を表します。同図からは、年齢が上がれば、平均所得も上昇する傾向があることがわかります。

2.3.1 条件付き平均値の限界

説明変数の数が多い場合、大量の条件付き平均値の数が計算されてしまい、理解が困難になります。例えば、年齢・性別・学位について平均賃金を計算した場合、40 個の組み合わせが発生します。データの事例数に比べれば、かなり少なくはなりますが、それでも理解には時間がかかるでしょう。

2.4 線型モデルへの要約

大量の変数間の関係性を要約するために、線形モデルを適用する方法が広く用いられます。例えば、賃金と年齢・学位・性別の関係を次のモデルで表すことを目指します。

$$\beta_0 + \beta_1 \times age + \beta_2 \times degree + \beta_3 \times gender \tag{2.2}$$

ただし degree と gender の値は、数字ではなく、文字です。線型モデルに導入する際には、何らかの数字に変換する必要があります。

ここでは bachelor と gender をダミー変数、degreebachelor と genderfemale、に変換します。ダミー変数は、以下のように定義されます。

- degreebachelor: degree が bachelor だった場合は 1、その他の場合は 0
- genderfemale: gender が female だった場合は 1、その他の場合は 0

$\beta_0, \beta_1, \beta_2, \beta_3$ はパラメタと呼ばれ、モデルがデータに極力一致するように選ばれます。このようにデータに合うように設定したモデルの特徴を調べることで、データの特徴を大雑把に理解することができます。

2.4.1 OLS

モデルをデータに適合される典型的な方法は OLS (最小二乗法) です。具体的な計算方法を紹介する前に、R での実行方法から紹介します。R では、lm 関数を用いることで、OLS

Table2.3

```
lm(earnings ~ age + degree + gender, CPSSW04)
```

```
Call:
lm(formula = earnings ~ age + degree + gender, data = CPSSW04)

Coefficients:
   (Intercept)             age  degreebachelor    genderfemale
        1.8838          0.4392          6.8651          -3.1579
```

を容易に実行できます。例えば、以下のコードで、Equation 2.2 を CPSSW04 に適合させることができます。

各説明変数名の下の数値が、パラメタの計算結果です。Equation 2.2 のようなシンプルな線型モデルにおいて、パラメタは各説明変数と結果変数である earnings の「モデル上での関係性」を表します。推定結果から、年齢や大学卒は賃金と正の関係があり、男性に比べて女性の賃金が低い傾向にあることがわかります。

2.4.2　モデルの定式化

線型モデルに要約することで、変数間の関係性を「ざっくり」把握することができます。注意が必要なのは、以上の方法では、どのようなモデルをデータに適合させるのかは、分析者が決定する必要がある点です。

どのようなモデルが適しているのかは、研究目標により異なります。「人間がデータの大雑把な特徴を理解する」ことを目標にするのであれば、Equation 2.2 のような、シンプルなモデルが有力です。このような大雑把な理解を目標に設定されたモデルは、**記述モデル**と呼ばれます[*1]。

[*1] データ分析では、記述を目的としないモデルの推定を目指すことがあります。例えば結果変数の値の予測

2.4.3 OLS の計算方法

OLS の具体的な計算方法を紹介します。基本的なアイディアは、データとモデルの乖離度合いを測定する指標である、**平均二乗誤差**を極力小さくするようにパラメタを算出するというものです。

例えば、age のみを説明変数として、earnings について、以下の線形モデルを設定します。

$$\beta_0 + \beta_1 \times age$$

ある事例 i についての二乗誤差は以下のように定義されます。

$$\text{事例}i\text{の二乗誤差} = (\text{事例}i\text{の所得} - \underbrace{(\beta_0 + \beta_1 \times \text{事例}i\text{の}age)}_{\text{モデルが算出する賃金}})^2$$

OLS は全ての事例について、二乗誤差の平均値 (平均二乗誤差) を最小にするようにパラメタを算出します。

平均値と同様に、OLS も別の計算方法があります。この方法は、分布と条件付き平均値のみを用いて計算できます。以下の条件付き平均値についての二乗誤差を最小化するようにパラメタの値を算出します。

$$\begin{aligned}
&(\text{25才の平均所得} - (\beta_0 + \beta_1 \times 25))^2 \times \text{25才の割合} \\
&+(\text{26才の平均所得} - (\beta_0 + \beta_1 \times 26))^2 \times \text{26才の割合} \\
&+...
\end{aligned}$$

こちらの計算方法では、パラメタを「モデルが条件付き平均値に極力適合するように決定」するので、算出されたモデルは条件付き母平均のモデルであると解釈できます。この解釈は、社会データの分析において現実的であり、多くの教科書で採用されています (Angrist and Pischke 2009; Aronow and Miller 2019; Chernozhukov et al. 2025)。

を目指すモデルは、予測モデルと呼ばれます。予想モデルについては、分析者がモデルの詳細な構造を理解することを**主たる目的としないため**、より複雑なモデルが設定されます。

2.4.4　線型モデルの限界

線型モデルは、あくまで関係性を単純化するため、実際の条件付き平均値を完全には反映しません。

例えば earnings について、

$$\beta_0 + \beta_1 \times age$$

を OLS で当てはめてみます。実際の条件付き平均値とモデルの関係性は、以下でした。

```
CPSSW04 |>
  ggplot(
    aes(
      x = age,
      y = earnings
    )
  ) +
  stat_summary(
    aes(
      color = "条件付き平均値"
    ),
    geom = "point"
  ) +
  geom_smooth(
    aes(
      color = "モデル"
    ),
    method = "lm",
    se = FALSE
  )
```

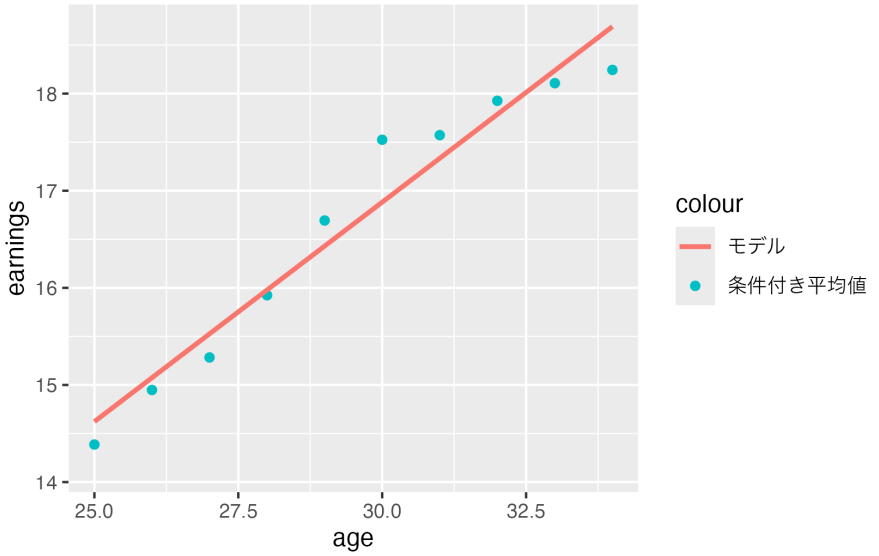


Figure2.1

各点は年齢ごとの条件付き平均値、直線が OLS で推定された線型モデルとなります。

線形モデルは、平均賃金と年齢が概ね右上がりの関係 (高年齢は賃金が高い) を捉えることができています。ただし実際の平均賃金と年齢の関係性は、モデルのように" 一直線" ではありません。線形モデルは、この特徴を捉えることには失敗しています。

2.5 モデルの複雑化

データ上の平均値とモデルの乖離は、推定するモデルを複雑化することで、簡単に削減できます。例えば以下のモデルを、OLS でデータに当てはめてみます。

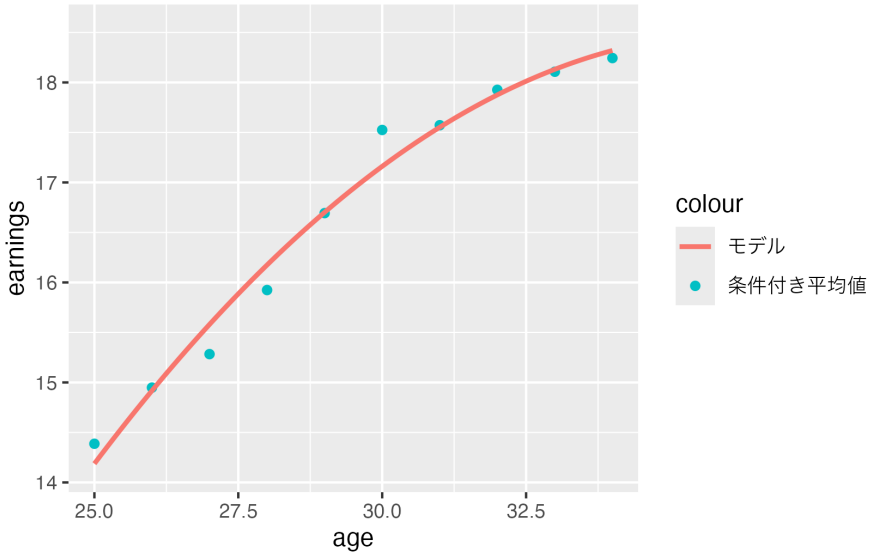
$$\beta_0 + \beta_1 \times age + \beta_2 \times age^2$$

$\beta_2 \times age^2$ は高次項と呼ばれます。このような項を加えても、OLS を用いればパラメータ

の推定は可能です。

下図は、この 2 次モデルをデータに当てはめた結果を示しています。直線ではなく、年齢に対して曲線的な関係が推定されており、データ上の条件付き平均値により近づいていることがわかります。

```
CPSSW04 |>
  ggplot(
    aes(
      x = age,
      y = earnings
    )
  ) +
  stat_summary(
    aes(
      color = "条件付き平均値"
    ),
    geom = "point"
  ) +
  geom_smooth(
    aes(
      color = "モデル"
    ),
    method = "lm",
    se = FALSE,
    formula = y ~ poly(x,2)
  )
```



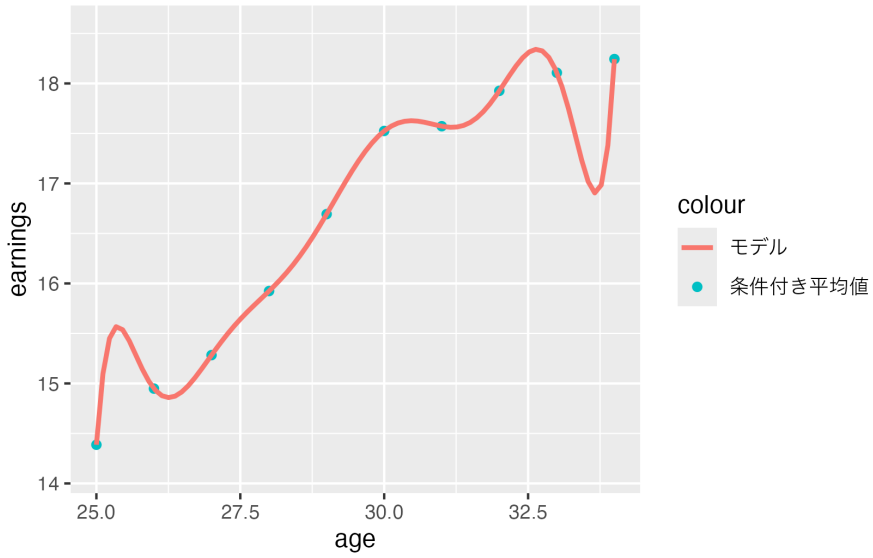
では、さらにモデルを複雑にするとどうなるでしょうか？ 次に、9 次までの高次項を含む以下のモデルを考えてみます。

$$\beta_0 + \beta_1 \times age + .. + \beta_9 \times age^9$$

このような高次の多項式モデルを用いると、モデルの予測値はデータ上の平均値とほぼ完全に一致するようになります。

```
CPSSW04 |>
  ggplot(
    aes(
      x = age,
      y = earnings
    )
  ) +
```

```
  stat_summary(
    aes(
      color = "条件付き平均値"
    ),
    geom = "point"
  ) +
  geom_smooth(
    aes(
      color = "モデル"
    ),
    method = "lm",
    se = FALSE,
    formula = y ~ poly(x,9)
  )
```



このように、線形回帰モデルは非常に柔軟な枠組みであり、どれほど複雑な形であっても、モデルが”β の足し算” である限り、データに当てはめること自体は可能です。

複雑なモデルには、以下のような魅力があります。

- データとの適合度が非常に高い
- 現実の社会や市場は複雑であるため、モデルも複雑な方が良いと感じられる

しかしながら、複雑すぎるモデルにはいくつかの問題点があります。

- 過学習（オーバーフィッティング）により、モデルの推定精度が悪化する（詳しくは Section 3.5 を参照）
- モデルの構造が複雑すぎると、人間がその意味を理解しにくくなる（詳しくは Chapter 4 を参照）

そのため、モデルの複雑さは、データの量や研究目標に応じて慎重に調整する必要があります。具体的な研究目標とそれに応じた OLS の活用方法は、Chapter 4 で紹介します。

第 3 章

母集団の要約と推定

平均値や OLS の結果は、計算に用いるデータによって異なります。例えば、1947 年に実施された第 1 回目の労働力調査と 2025 年に行われた調査では、平均賃金が大きく異なるでしょう。これは 1947 年と 2025 年では、社会状況が大きく異なるので、当然の結果と考えられます。

では全く同じ社会を対象とした調査を、複数の研究チームが行った場合に、同じ平均賃金が算出されるでしょうか？ もし同じ結果が得られないのでは、分析結果には、「客観性がなく信用できない」とも考えられます。このような場合、分析結果をどのように受け取れば良いでしょうか？

以上の問題は、推定問題と呼ばれ、データ分析における中心的な課題の一つです。

❗ 到達目標

- 推定対象と推定結果を明確に区別する
  - 推定対象と推定結果を定義するための概念: 母集団とサンプルングを理解する
- 推定対象を推論するツール、信頼区間を理解する

3.1　推定問題の枠組み

3.1.1　頻度論

推定問題を考える土台となる枠組みとしては、頻度論とベイズ論が有名です。本ページでは、頻度論に基づく議論を紹介します。

まずは、以下の**思考実験**を考えてください。

> ❗ Important 1: 思考実験
>
> 日本全国に、「2025 年の日本の労働市場」を分析する研究チームが、大量に組織されました。各チームは同じ手順に基づき、平均賃金を計算します。ただしデータの収集は、各チームが**独立**して行います。
> **果たして、各チームは同じ結論にたどり着くでしょうか？**

もし全てのチームが同じデータと同じコードを使えば、コーディング・ミスなどがない限り同じ結果になるはずです。ところが、各チームが電話・ネット調査などでそれぞれが独自にデータを集めた場合、調査対象の対象となった回答者の違いから、得られるデータが異なり、結果も変わる可能性があります。

この思考実験の結果を順序立てて想像するために、いくつかの理論的な概念を導入します。

3.1.2　推定結果と推定目標

次の 2 つを区別することが重要です:

- **推定結果**（estimate）：データから実際に計算される値
- **推定目標**（parameter of interest/estimand）：推定の対象となる “真の値”

たとえば、CPSSW04 データから計算される平均所得は、あくまで「推定結果」であり、「全米の平均所得」という推定目標に近づくことを期待して計算されます。

推定結果と推定目標の関係性を明確にするために、**母集団**と**サンプルング**という概念を導入します。

3.2 母集団とサンプルリング

3.2.1 母集団

母集団 (Population) とは、推定の対象となる集団のことです。Important 1 における母集団は「2025 年の労働者」、CPSSW04 では、「2004 年のアメリカの全世帯」が妥当な母集団でしょう。

3.2.1.1 母分布

母分布は、母集団における変数の分布です。本ページでは、データに含まれる変数について、母集団全員を調査し、計算された分布であるとイメージしてください。

例えば CPSSW04 の母集団である 2004 年のアメリカの全世帯が観察できれば、同年の特定の [収入、学位、性別、年齢] が、全人口に占める割合が計算できるはずです。

以下では、「分析者は母集団を直接観察できず、**母分布を実際に計算することは不可能である**」、という状況を想定します。そして、直接知ることができない母分布やその特徴を、データから推定することを試みます。

3.2.2 推定目標の定義

本ページでは、母分布そのものではなく、母分布の特徴を推定する方法を紹介します。推定の対象となる母分布の特徴を、推定目標と呼びます。代表的なものとしては、母分布から計算された**平均値**および **OLS の結果**があります。

3.2.2.1 母平均

データにおける分布と同様に、母分布からも平均値や条件付き平均値を計算できます。このような母集団における平均値を、母平均と呼びます。

例えば、母集団における 25 才の平均所得は以下のように計算できます。

$$
\begin{aligned}
&25\text{才の条件付き母平均} \\
&= 1 \times 25\text{才における「}earnings\text{が1」の条件付き母分布} \\
&+2 \times 25\text{才における「}earnings\text{が2」の条件付き母分布} \\
&+...
\end{aligned}
$$

3.2.2.2　Population OLS

本スライドでは、母分布から計算される OLS の結果を推定目標とします。このような推定目標を母集団における OLS 推定値 (Population OLS) と呼びます。

例えば、賃金と年齢の関係性を捉えるために、以下の母集団における平均二乗誤差を最小化する線型モデル $(\beta_0 + \beta_1 \times age)$ を Population OLS として定義できます

$$
\begin{aligned}
&(25\text{才の母平均} - (\beta_0 + \beta_1 \times 25))^2 \times 25\text{才の母分布} \\
&+(26\text{才の母平均} - (\beta_0 + \beta_1 \times 26))^2 \times 26\text{才の母分布} \\
&+...
\end{aligned}
$$

これらはすべて、実際には観察できない仮想的な計算結果です。この結果をデータから推定することを目指します。

3.2.3　サンプリング

データは、何らかの方法で選ばれた事例の集まりであると想定します。この事例を選ぶ過程を、**サンプルング** (sampling) と呼びます。

母集団について推測するためには、データと母集団の関係について何らかの仮定を置く必要があります。特に、母集団の事例をどれだけ偏りなくサンプリングできているのかは、分析の信頼性に大きく影響します。

最も重要な仮定は、ランダムサンプリングです。

> ❗ Important 2: ランダムサンプルングの仮定
>
> - データの各事例は、母集団から個別にランダムに選ばれている

本ページでは、分析に用いるデータはランダムサンプルングを満たすことを想定します。

3.2.4 推定結果と推定方法

推定結果（estimate）とは、データから計算される値のことです。この推定結果は、推定対象（母集団上での真の値）に近い値であることが期待されます。

推定結果を計算する手順のことを、**推定方法（estimator）**と呼びます。

3.2.5 まとめ

本節で登場した重要な概念と仮定を整理すると、次のようになります。

> ❗ 母集団、サンプルング、推定対象、推定結果
>
> - **母集団**: 推定対象となる集団
> - **推定対象**: 母集団上で行いたい仮想的な計算 (母集団上で行いたい計算)
> - **サンプルング**: 母集団から事例を収集する手順
> - **ランダムサンプルングの仮定** (Important 2) : 事例は、母集団からランダムに選ばれている
> - **データ**: サンプリングされた事例の集団
> - **推定結果**: データから計算される結果 (推定対象に近い値であることが期待される)
> - **推定方法**: データから推定結果を計算する具体的な計算方法

これらの関係は、以下の図にまとめられます：

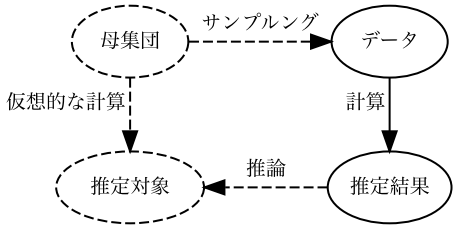


Figure3.1: “推定問題”

図中の実線は、データの分析者が実際に観察・操作できる要素を、点線は想像上の操作や概念を表しています。例えば、母集団から推定対象を計算する作業を、分析者が実際に行うことはできません。なぜならば、母集団を直接観察することができないためです。

> ❗ ポイント
>
> ここまでの枠組みを用いると、Important 1 の思考実験の結果は以下のように整理できます。
>
> - データの事例はランダムに選ばれているため、データから得られる推定結果は研究チームによって異なる

3.3　推定方法

母平均や Population OLS は、以下の一般的な推定方法が利用できます[*1]。

> ❗ Important 3: 母平均や Popualtion OLS の推定
>
> 1. **母集団上での仮想的な計算結果** として、推定対象を定義する
> 2. **同じ計算をデータ上で行った結果**を、推定結果とする

[*1] アナログ原理やプラグイン原理として呼ばれています。

この推定方法は、以下の性質から正当化できます。

> **!** Important 4: 一致性
>
> もしデータに代表性があり、かつ事例数が無限大であれば、データから算出した OLS の結果と Population OLS は一致する。

一致性からは、データが増えれば増えるほど、推定結果は推定目標に近づくことが期待できます。

ただし、この議論は「母集団において平均値や OLS の結果が計算できる」ということを、前提にしていることに注意してください。例えば、母分布が「特殊な」場合、母平均が無限大になり、計算できない場合があります[*2]。また Population OLS については、母集団において多重共線性が存在する場合、計算は不可能です。そのため、多重共線性のない定式化を用いて、Population OLS を定義する必要があります。

3.3.1 一致性の限界

一致性は、推定方法 Important 3 を正当化する重要な性質です。しかしながら、実際のデータ分析においては、それほど実用的な性質ではありません。なぜならば、Population OLS とデータ上での OLS が一致するためには、**無限大の事例数**が必要となるからです。いうまでもなく無限大の事例数を持つデータは存在しません。言い換えると、実際のデータ分析では、推定結果と推定目標は” 乖離している” と想定すべきです。

具体的な例から考えてみます。

```
library(tidyverse)

data("CPSSW04", package = "AER")

lm(earnings ~ age, CPSSW04)
```

[*2] 母平均が計算できないケースとしては、母分布がコーシー分布である場合などが有名です。

```
Call:
lm(formula = earnings ~ age, data = CPSSW04)

Coefficients:
(Intercept)          age
     3.3242       0.4519
```

以上の結果は、データ上での OLS では、age のパラメタは 0.4519 であることが確認できます。一致性から、もし CPSSW04 がランダムサンプルングの仮定を満たし、事例数が無限大であれば、「**Population OLS** における age のパラメタも 0.4519 である」という結論は必ず正しいものとなります。ところが実際の事例数は、7986 であり、無限ではありません。このため「**Population OLS** における age のパラメタも 0.4519 である」は、ほぼほぼ間違った結論となります。

! ポイント

ここまでの議論から、Important 1 の思考実験の結果は以下のように整理できます。

- 各研究チームは、ランダムサンプリングにより、データを集めたとする
  - データの事例数が無限大あれば、データから得られる OLS の結果は、常に Population OLS と一致する
    - データから得られる推定結果は、どの研究チームでも同じ
  - 現実的な事例数のもとでは、データから得られる OLS の結果は、常に Population OLS から乖離
    - データから得られる推定結果は、研究チームよって異なる

3.4 信頼区間

3.4.1 統計的推論

ほぼほぼ間違った結論ではなく、正しい結論を示すことは可能でしょうか？ データ分析においては、**確実に正しい**結論を示すことは、事実上不可能です。このため多くのデータ分析では、**ほぼほぼ正しい**結論を示すことを目指します。このような**ほぼほぼ正しい**結論を示すプロセスは、**統計的推論**と呼ばれます。

3.4.2 信頼区間

統計的推論に活用できるツールは、さまざまなものがあります。代表的なものとして、信頼区間を紹介します。

信頼区間

- 推定対象を一定の確率で含むと考えられる、データから計算される区間
  s- 推定対象を含む確率は、信頼確率と呼ばれ、研究者が指定する

R において、信頼区間を計算する方法は複数存在します。例えば estimatr パッケージの lm_robust 関数を用いれば、以下の仮定のもとで、信頼区間を計算します。

信頼区間計算の前提条件

- データはランダムサンプルングされている
- 事例数は十分にある[*3]
- 極端なハズレ値がない

[*3] どのくらいあれば十分なのかは、難しい問題です。ただし多くの教科書では、200 事例以上が基準とされています。

例えば、以下のコードから、reform の平均値について信頼区間を計算できます。level には、信頼水準 (信頼区間が推定対象を含まむ確率) を指定します。省略すれば、自動的に 0.95 が指定され、95% 信頼区間が計算されます。

Table3.1

```
estimatr::lm_robust(earnings ~ age, CPSSW04)
```

```
             Estimate Std. Error   t value     Pr(>|t|)  CI Lower  CI Upper
(Intercept) 3.3241841  0.9657404  3.442109 5.801562e-
04 1.4310807 5.2172874
age         0.4519313  0.0329688 13.707847 2.737848e-
42 0.3873039 0.5165588
              DF
(Intercept) 7984
age         7984
```

以上の結果から、「Population OLS における age は、概ね 0.387 ～ 0.517 である」という主張を行うことができます。

> ❗ ポイント
>
> ここまでの議論から、Important 1 の思考実験の結果は以下のように整理できます。
>
> - 各研究チームは、ランダムサンプリングにより、十分な事例数があるデータを集め, 95% 信頼区間を計算したとする
>   - 全研究チームの 95% は、推定対象を含む区間を算出する

3.5 複雑すぎるモデルの問題点

モデルが複雑すぎると、推定において深刻な問題が生じることがあります。複雑なモデルのパラメータを高い精度で推定するためには、十分な数の観測事例が必要になるためです。このためデータ数が限られている場合、複雑なモデルを推定すると、推定精度が大幅に低下し、ミスリードな推定結果を導くおそれがあります。

ここでは、単純な数値例を用いてこの問題を確認してみます。

数値例

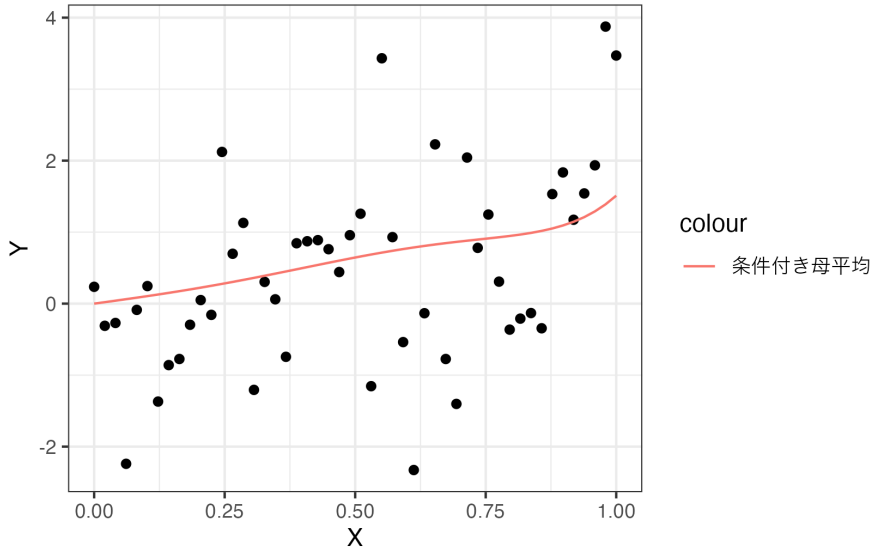
以下のような複雑な条件付き母平均を想定します[*4]。

$$E[Y \mid X] = X + 0.1 \times X^2 + 2 \times X^3 + 0.01 \times X^4 + 5 \times x^5$$
$$0.1 \times x^6 + 3 \times x^7 + 0.1 \times x^8 + 0.1 \times x^9 + 0.1 \times x^{10} \quad (3.1)$$

条件付き母平均とこの想定に基づいてランダム・サンプルングされた 50 事例のデータを図示すると、以下のようになります。

[*4] 個々の Y と X の値は、以下で決定されます: $X = 0$ から 1 までの一様分布、

$$Y = E[Y \mid X] + \underbrace{u}_{\text{平均0,分散1の正規分布}}$$

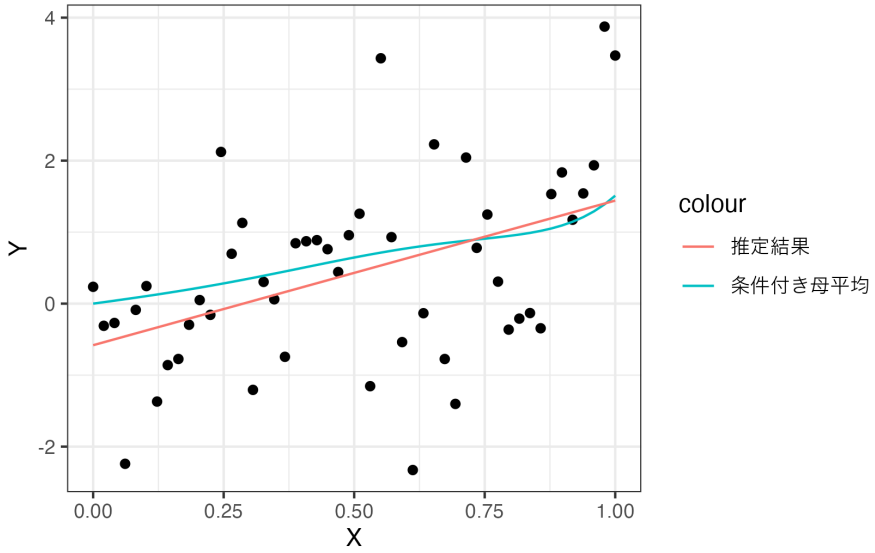


赤線が条件付き母平均、黒点がデータを表します。

まずは、単純な線型モデル $(\beta_0 + \beta_1 \times X)$ を推定してみます。

```
fixest::feols(
  Y ~ X,
  data
)
```

この推定結果と条件付き母平均を比較すると、以下のようになります。

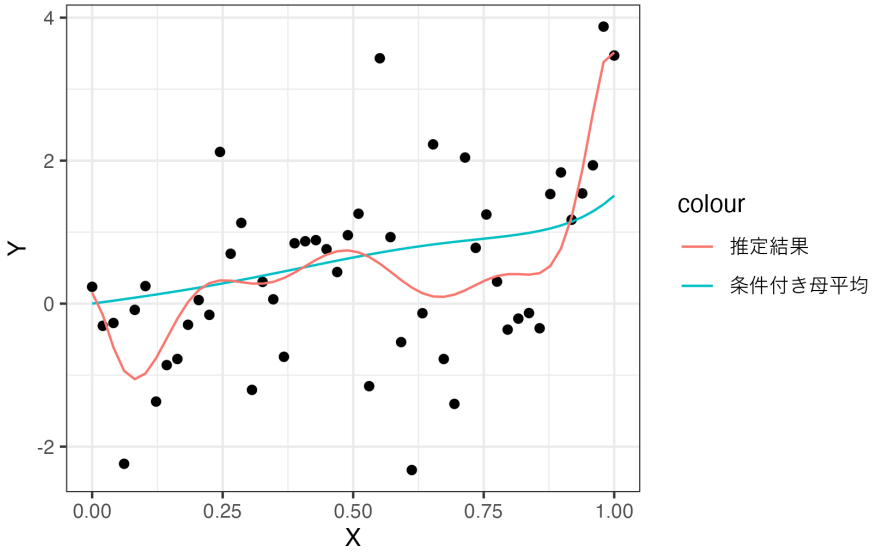


この図からわかるように、単純なモデルでは複雑な母平均を捉えることができず、推定結果と母平均に乖離が生じています。

次に、Equation 3.1 のような複雑な条件付き母平均を想定し、以下のような複雑なモデルを推定してみます。

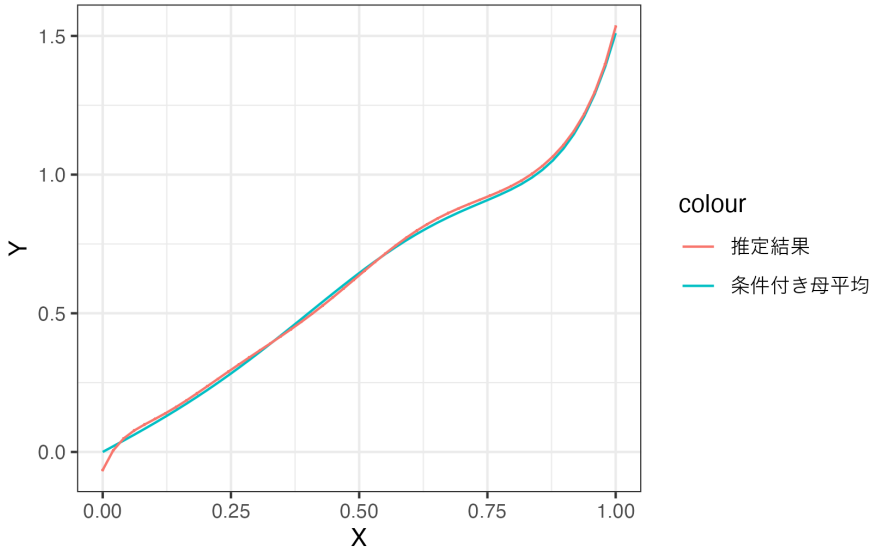
```
fixest::feols(
  Y ~ poly(X,10), # X の 10 乗までがモデルに導入されます。
  data
)
```

この推定結果と条件付き母平均を比較すると、以下のようになります。



複雑なモデルの方が、条件付き母平均に近づくように思えるかもしれませんが、実際には母平均との乖離がむしろ拡大していることがわかります。これは、少ないデータ（50 事例）で複雑なモデルを推定したため、パラメータの推定精度が著しく低下したことを示しています。

では、同じ複雑なモデルを、十分なデータ（5 万事例）で推定した場合はどうなるでしょうか。



この場合、推定結果は条件付き母平均とほぼ一致しており、十分なデータがあれば、複雑なモデルでも OLS によって高精度な推定が可能であることが確認できます。

この数値例からわかるように、複雑なモデルを用いる際には、十分なデータがあるかどうかを慎重に検討する必要があります。

第 4 章

OLS の応用と課題

ここまで、あるモデルを**どのように推定するか**を紹介してきました。しかし、**どのようなモデルを推定すべきか**は、研究の目的によって異なります。本章では、望ましいモデルの定式化について、目的別に整理していきます。またそれぞれの研究目的について、OLS を用いる際の注意点と OLS 以外の手法も紹介します。

4.1　ナイーブな見方: 完璧なモデルの推定

データ分析を学び始めたばかりの頃、「この手法を使えば社会の真理を明らかにできるのではないか」と期待してしまうことがあります。言い換えれば、結果変数に対して「完璧でシンプルなモデル」を推定できると考えてしまうかもしれません。

たとえば、以下のような回帰モデルを用いて、賃金を正確に予測できると期待することがあります。

しかしながら、このような期待は、少なくとも社会や市場のデータ分析においては、現実的ではありません。以下の点に注意が必要です。

- OLS によって推定されるのは、あくまで「条件付き母平均」のモデルです。つまり、特定の条件（たとえば学歴・性別・年齢）を満たす人々の平均的な賃金を推定しているにすぎません。現実の社会や市場では、同じ条件を持つ人々の間でも賃金

Table4.1

```
data("CPSSW04", package = "AER") # データの読み込み

fixest::feols(earnings ~ degree + age + gender, CPSSW04)
```

```
OLS estimation, Dep. Var.: earnings
Observations: 7,986
Standard-errors: IID
                Estimate Std. Error   t value  Pr(>|t|)
(Intercept)     1.883797   0.920292   2.04696  0.040695 *
degreebachelor  6.865150   0.178369  38.48856 < 2.2e-16 ***
age             0.439204   0.030529  14.38664 < 2.2e-16 ***
genderfemale   -3.157864   0.180365 -17.50821 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 7.88234   Adj. R2: 0.189694
```

には大きなばらつきがあります。たとえば、同じ学歴・性別・年齢であっても、職種や勤務地、経験年数などによって賃金は大きく異なるのが普通です。したがって、モデルが「真の賃金」を正確に予測することは不可能です。

- 上記のような単純なモデルでは、現実の平均賃金の構造を十分に表現することはできません。現実の平均賃金と学歴、性別、年齢の関係性は、より複雑であると予想されます。

以上の理由から、OLS によって「完璧でシンプルなモデル」を推定しようとする試みは、通常うまくいきません。そのため、分析の目的に応じて、適切なモデルを構築することが現実的なアプローチとなります。

4.2 全体の記述

記述研究とは、データに含まれる情報をもとに、複数の変数の間にどのような関係が存在するのかを明らかにすることを目的とする研究です。

このような目的に対して、OLS は非常に有効な分析手法とされています。OLS は、単純な線形モデルの推定を得意としています。Table 4.1 のような単純な線形モデルは構造が単純であるため、ある程度データ分析に慣れていれば、どの変数がどのように影響しているのかを人間が理解しやすい形で示すことができます。例えば、大学卒のパラメタは正なので、平均賃金が高い傾向、女性は平均賃金が低い傾向、年齢と平均賃金は正の関係性が読み取れます。

4.2.1 OLS の注意点

しかし、OLS を用いる際には注意すべき点もあります。特に重要なのは、得られた関係性の要約が、どのようなモデルを推定したのか、すなわちモデルの定式化に大きく依存するという点です。たとえば、どの変数を説明変数として選ぶか、自条項などを加えるかなどの選択によって、分析結果が大きく変わる可能性があります。そのため、モデルを定式化する段階で、定式化が最終的な解釈や結論に強く影響を与えることを理解しておく必要があります。

4.2.2 他の手法

本ページでは、OLS は条件付き母平均の線型モデルを推定する手法として紹介しました。言い換えると、条件付き母平均の記述の手法です。

記述分析の目標は、条件付き母平均のみではありません。たとえば、母分布そのものの記述を目指す場合には、クラスタリングモデル[*1]などが有効です。このような分布のモデルを推定する手法としては、最尤推定やベイズ推定法が活用されます。

[*1] Gormley, Murphy, and Raftery (2023)

4.3　予測

予測研究とは、推定されたモデルを用いて、目的変数 Y の値や未知を予測することを主な目的とする研究です。たとえば、Table 4.1 で推定されたモデルを用いて、ある個人の属性情報（年齢、学歴）をもとに、その人の賃金を予測する場合などが該当します。

予測研究と比較研究との大きな違いは、モデルの中身、すなわち各パラメータの意味や解釈が重視されない点にあります。比較研究では、変数間の関係性を人間が把握することが目的となるため、モデルの理解やパラメータの解釈が重要です。一方、予測研究では、最終的な予測の精度こそが最も重要な評価基準となります。

このため、予測研究では、人間が予測モデルの構造を完全に理解できなくても、大きな問題とはされません。むしろ、より高い予測精度を実現するために、複雑なモデルが積極的に用いられる傾向があります。たとえば、以下のような二乗項と交差項 (変数同士の掛け算) を含むモデルが考えられます。

$$
\begin{aligned}
earningsの予測モデル = \beta_0 \\
+\beta_1 \times degreebachelor + \beta_2 \times age + \beta_3 \times genderfemale \\
+\beta_5 \times age^2 \\
+\beta_6 \times degreebachelor \times age + \beta_7 \times degreebachelor \times genderfemale \\
+\beta_7 \times age \times genderfemale
\end{aligned}
$$

このような複雑なモデルであっても、OLS によって推定することが可能です。

```r
model <- fixest::feols(
  earnings ~ degree + age + gender +
    age^2 +
    degree:age + degree:gender + age:gender,
  CPSSW04)

coef(model)
```

```
                (Intercept)              degreebachelor
               -21.60207344                -5.33505137
                        age               genderfemale
                2.11156693                5.10734867
                   I(age^2)         degreebachelor:age
               -0.02946533                 0.41195443
degreebachelor:genderfemale           age:genderfemale
               -0.24754070                -0.27156417
```

4.3.1 OLS の注意点

予測分析においても、モデルを複雑化することの弊害は存在します。それは、パラメタの推定精度の悪化に伴う、予測精度の低下です。予測性能を確保するためには、モデルを複雑にしすぎないことが必要です。

ただし、特に説明変数が多い場合、研究者がモデルの複雑性を適切にコントロールすることは、事実上不可能です。

4.3.2　他の手法

近年では、より柔軟にデータを活用する手法が注目されています。代表的なものとして、機械学習分野における教師付き学習[*2]があり、経済学の分野でもその応用が急速に進んでいます。これらの手法は、複雑なモデルであっても予測精度を確保するためのさまざまな工夫が導入されている点が特徴です。

4.4　特定の特徴の記述

OLS がよく応用されてきた研究課題の一つに、特定の変数間の関係性を明らかにする分析があります。たとえば、賃金（earnings）と学歴（degree）の関係を調べる研究がその例です。

このような課題に対しては、以下のようなシンプルな線形モデルを推定することで、基本的な関係性を把握することができます。

```
model <- fixest::feols(earnings ~ degree, CPSSW04)

confint(model)[2,] # degreebachelor についての信頼区間を抽出
```

```
                 2.5 %   97.5 %
degreebachelor 6.138989 6.855967
```

この推定結果から、大学卒の方が、概ね 6.14 ～ 6.86 ほど平均賃金が高いと考えられます。

さらに、OLS を用いることで、賃金と学歴の関係について、より詳細な特徴を捉えることも可能です。たとえば、「もし大学卒と非大学卒の間で、性別や年齢の分布に違いがなかったとしたら、賃金格差はどうなるか？」という問いに答えることができます。

[*2] James et al. (2021)

Table4.2

```
model <- fixest::feols(
  earnings ~ degree +
    age + gender +
    age^2 + age:gender, # gender と age について交差項と二乗項を導入
  CPSSW04)

confint(model)[2,] # degreebachelor についての信頼区間
```

```
                2.5 %   97.5 %
degreebachelor 6.4815 7.180926
```

実際のデータでは、大学卒と非大学卒の間で性別や年齢の分布に差があります。こうした差を統計的に調整したうえで賃金を比較するには、直接の関心ではない説明変数（この場合は age や gender）について、より柔軟に定式化したモデルを用いることが有効です。

以下は、その一例です。

この推定結果から、大学卒の人の平均賃金は、概ね 6.49 ～ 7.18 ほど高いという結果が得られ、単純なモデルよりも格差がやや広がっていることがわかります。

4.4.1 OLS の注意点

結果の解釈

このように推定された差が、社会のどのような特徴を捉えているのかは、以前として不明確です。例えば、degreebachelor の信頼区間は 6.49 ～ 7.18 であったとしても、「大学進学が賃金を平均的に増加させる因果的効果を持つ」と結論づけることはできません。

その理由は、degree 間の賃金格差が、age や gender 以外の要因によって生じている可能性があるからです。

このように、データから得られた推定結果をもとに因果効果をどのように推論するかという問題は、「統計的因果推論」と呼ばれる分野で活発に研究されています[*3]。

推定の信頼性

Table 4.2 では、関心のある変数 degree と、その他の変数 age や gender の交差項は導入していません。このような場合、degreebachelor の推定結果の解釈が不明瞭になることがあります。

そのため、degree と他の変数との交差項を導入し、それらの結果を「集計」するというアプローチが推奨されます[*4]。このような推定は、marginaleffects パッケージを用いることで簡単に実装できます[*5]。

たとえば、以下のように実行します。

```
model <- fixest::feols(
  earnings ~ degree +
    degree:(age + gender +
              age^2 + age:gender) + # degree との交差項
    age + gender +
    age^2 + age:gender, # gender と age について交差項と二乗項を導入
  CPSSW04)

marginaleffects::avg_slopes(
  model,
  variables = "degree"
)
```

```
 Estimate Std. Error    z Pr(>|z|)      S 2.5 % 97.5 %
```

[*3] Chernozhukov et al. (2025), Wager (2024) , Ding (2024) (ドラフト版: Ding (2023))
[*4] Chattopadhyay and Zubizarreta (2023)
[*5] 詳細は、パッケージのホームページ (https://marginaleffects.com/) を参照ください。

```
     6.82      0.178 38.3   <0.001 1064.6  6.47   7.17

Term: degree
Type: response
Comparison: bachelor - highschool
```

この結果から、gender と age の分布の違いを統計的に解消した場合、大学卒の人の平均賃金は概ね 6.47 ～ 7.17 高いと推定されました。

4.4.2 他の手法

OLS を用いた比較には、いくつかの問題点が指摘され、それを解消できる手法も開発されてます。例えば、傾向スコアやエントロピーウェイトなどを活用する方法 (Hainmueller 2012) が有名です。

さらに機械学習を活用した信頼性の改善方法も、近年確立されつつあります。この方法は機械学習の持つ問題点 (信頼区間の計算が難しいなど) を補う仕組みを導入することで、モデルの定式化への推定結果の依存度を下げつつ、推定誤差を考慮した分析が可能になっています[*6]。

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.

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

Chernozhukov, Victor, Christian Hansen, Nathan Kallus, Martin Spindler, and Vasilis Syrgkanis. 2025. *Applied Causal Inference Powered by ML and AI*. https://causalml-book.org/.

Ding, Peng. 2023. "A First Course in Causal Inference." https://arxiv.org/abs/2305.18793.

[*6] Chernozhukov et al. (2025)

———. 2024. *A First Course in Causal Inference*. Chapman; Hall/CRC.

Gormley, Isobel Claire, Thomas Brendan Murphy, and Adrian E Raftery. 2023. "Model-Based Clustering." *Annual Review of Statistics and Its Application* 10 (1): 573–95.

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.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning: With Applications in r*. 2nd ed. Springer Texts in Statistics. Springer. https://www.statlearning.com/.

Wager, Stefan. 2024. *Causal Inference: A Statistical Learning Approach*. Technical report, Stanford University. https://web.stanford.edu/~swager/causal_inf_book.pdf.