計量経済学 2025

講義ノート

高知工科大学　経済マネジメント学群

藤井大和(1270526)

目次

線形回帰

アウトカムの平均値が、説明変数の線形関数で定義される値の変化に応じて変化する

線形関数

加法性：$f(x+y)=f(x)+f(y)$

斉次性：$f(kx)=kf(x)$

を満たす必要がある。つまり直線になる関数のこと。

応答変数（従属変数）

結果となる変数

説明変数（独立変数）

結果に影響を与える要因

つまり応答変数を説明変数に回帰する

単回帰分析

説明変数が1つ

重回帰分析

説明変数が2つ以上

回帰分析の基礎

図示する：散布図

統計量を求める：相関係数

相関係数が同じでも、グラフの傾きは異なる可能性がある

傾き：説明変数が与える応答変数への影響の大きさ

→相関係数がわかっても、予測ができない

予測できるようにするのが回帰分析

回帰直線

応答変数と説明変数の関係を示す直線

傾きがわかる（1単位増えるとどれだけ増えるかがわかる）→予測可能

回帰分析には、1つの応答変数と1つ以上の説明変数が必要

説明変数をx、応答変数をyとすると

$$y = ax + b$$

a:傾き　b:切片

残差

回帰直線と実際の点は多くがずれている。

回帰直線と実際の点の間を残差（e）という

散布図上の点と、直線$y = ax + b$からのズレに分解

$$y_i = a + bx_i + e_i = \hat{y}_i + e_i$$

ただし、$i = 1,2, \dots, N$

$\hat{y}_i$：予測値

観測値＝予測値＋残差

→この残差を小さくしたい

最小二乗法（OLS）

残差の平均値は必ず 0 になる（プラスマイナス打ち消しあう）

残差の二乗の総和した合計（残方平方和）を最小化することで、残差の少ない観測値に近い直線を求める手法（二乗することでマイナスを打ち消し、その合計を最小化する）

単回帰分析

モデル1　説明変数がダミー変数の場合

衆議院議員総選挙での得票率を、衆議院議員経験の有無で説明する

応答変数：得票率（%）

説明変数：衆院議員経験がある（現職，元職）候補者は 1、その他は 0 のダミー変数

推定結果：得票率 ＝ 14 ＋ 31×議員経験 ＋ 残差

$\widehat{得票率} = 14 + 31 \times 議員経験$

モデル 1 の予測値：議員経験（0 または 1）が与えられたときの、得票率の平均値（期待値）

$\widehat{得票率} = 14 + 31 \times 議員経験$

$\widehat{議員経験がない候補者の得票率} = 14 + 31 \times 0 = 14$

$\widehat{議員経験がある候補者の得票率} = 14 + 31 \times 1 = 45$

回帰係数：議員経験がある候補者と議員経験がない候補者の平均得票率（予測値）の差

モデル2　説明変数が量的変数の場合

衆議院議員総選挙での得票率を、選挙費用の大きさで説明する

応答変数：得票率（%）

説明変数：選挙費用（測定単位：100 万円）

推定結果：$得票率 = 7.7 + 3.1 \times 選挙費用 + 残差$

回帰直線上の点：

- ✓ 選挙費用ごとに予測される得票率
- ✓ 候補者を選挙費用ごとにグループ分けしたときの、グループの平均得票率

$得票率 = 7.7 + 3.1 \times 選挙費用 + 誤差$

選挙費用の係数 3.1

- ✓ 選挙費用の値が 1 だけ異なる候補者を比べると、選挙費用が大きいほうが、平均すれば 3.1 ポイント高い得票率を得る
  - ➢ 選挙費用を 100 万円増やすと、得票率は 3.1 ポイント上がると期待される
  - ➢ 選挙費用を 1,000 万円増やすと、得票率は 31 ポイント上がると期待される

切片 7.7

- ✓ 「選挙費用 ＝ 0」の候補者の平均得票率を示す
- ✓ 選挙費用が 0 の候補者は存在しないので、この場合切片には意味がない

データの収集・クリーニング

データの入手法

長方形の形(CSV や Excel)で入手できるデータ

- ✓ 公的機関のウェブサイト
- ✓ 研究者や大学の website
- ✓ オープンデータアーカイブ

長方形の形ではない場合

- ✓ 手入力
- ✓ コピペ
- ✓ スクレイピングソフト(R でウェブスクレイピング)
- ✓ 図書館での CD-ROM やオンラインデータベース　書籍からの入力
- ✓ 販売されているタイプのデータ。高額なものなので、図書館が所蔵していないか、買ってもらえないかなどお願いする。

作る

- ✓ 調査や観察、実験によってデータを集める。ただし、データソースを含めてすべてを記述しなければならない。コーディング(仕分け)のルールを決めて、文書として記録する。

分析に適したデータの方式

最も一般的なのは、長方形データ(Rectangular Data)。いわゆる表形式データのこと。

- ✓ 各行が観測単位一つを表す
- ✓ 各列が 1 つの変数を表す
  - ➢ CSV ファイル
    - ✧ カンマ区切りのテキストファイル
    - ✧ MS Excel やメモ帳で編集可
    - ✧ ファイルサイズが小さく、共有がしやすい

Tidy Data

分析を円滑に行うために、データを使いやすい良い形で準備したい。

R の分析においては、Tidy Data 化が求められる。

Tidy Data:整然データ　データの構造と意味が一致しているデータのこと

Tidy Data4条件

1. 1 つの列は、1 つの変数を表す
2. 1 つの行は、1 つの観測を表す

3. 1 つの表は、1 つの観測単位（unit of observation）を表す
4. 1 つのセルは、1 つの値を表す

可読性は Tidy Data とは両立しないことが多い。しかし、PC が読み取りやすいデータは Tidy Data である。政府統計なども観測単位が違うデータが一つの表にまとまっているため、Tidy Data になっていない。

★Tidy data の特徴

列：変数

行：観測

セル：値

表（データセット）：1つの観測単位に基づいて集められた情報

データ分析：変数間の関係の意味を調べたい

R でプログラミングするときには、意味ではなく構造に頼る必要があるため

回帰分析における統計的推測Ⅰ　～仮説の立て方～

データの種類

データとは何か

調査や観察・実験などで集められた情報

- ✓　数量データ：数値化されたもの
- ✓　質的データ：数値化できないもの

我々は、観察の対象によって値が変わるもの（変数）に興味があり、動かない値（＝定数）には興味がない

変数

- ✓　数が一定でない＝変化する数
- ✓　様々な値をとる（つまり分布する）

変数の分類		カテゴリー間の異同	順序	差	比
質的変数	名義尺度	✔	-	-	-
	順序尺度	✔	✔	-	-
量的変数	間隔尺度	✔	✔	✔	-
	比率尺度	✔	✔	✔	✔

1. **名義尺度**
   通称：カテゴリ変数
   「違い」は量的ではなく、質的なもので、大小は比べられない　つまり識別のためだけに存在し、数字そのものに意味はない。
   Ex）ダミー変数
2. **順序尺度**
   対象間の順序（大小、長短、強弱）の情報を与える。
   差を比べることはできない。（差は均等とは限らない）
   つまり大小はわかるが、どれくらい大きいか小さいかが分からないデータのこと
   Ex）大学の成績（AA, A, B, C, F）、スポーツの順位
3. **間隔尺度**
   各順位間の差・距離を等しい単位で設定したもの
   ただし、対象間の比率を測ることはできない。つまり乗法除法ができない。
   Ex）摂氏温度と華氏温度　華氏が2倍になったから摂氏も二倍になるかといえば、そういうわけではない。
4. **比率尺度**
   対象間の「比」の情報を与える

絶対的な原点「0」が存在する
4尺度の中で情報量が最も多く、四則演算が可能な”普通の”数字データ
Ex)絶対温度　身長　体重　年収

理論と仮説

本質的には違いはない　理論は受け入れられた仮説にすぎない

- ✓ **理論仮説**
  原因と結果の関係についての抽象的な仮説
  Ex)学歴が高いほど政治に参加する
- ✓ **作業仮説**
  理論仮説から抜き出された特定の変数に関する具体的な仮説
  理論仮説から引き出される観察可能な予測について述べる
  理論仮説の検証のために、正しいと仮定した仮説として扱う
  Ex)学校へ通った年数が高いほど、国政選挙への投票率が高い
  Ex)大卒の人ほど、デモに参加しやすい

観察可能な作業仮説が多い理論仮説がより支持される(＝反証がたくさんされているから)

操作化(Operationalization)

理論仮説中の変数を、観察可能かつ計量可能な変数に置き換えること
Ex)人間の知性　→　知能テストの点数
Ex)国の経済力　→　一人当たりのGDP(国内総生産)

理論の操作化の例

理論「教育が政治参加を促進する」
↓操作化
「偏差値が高いほど投票率が高い」

操作化の方法は一つではない

理論仮説の変数にできる限り近いものを選ぶ
作業仮説が理論仮説から乖離すると、作業仮説を検証しても、理論を検証したと信じてもらえなくなる
観察可能な変数を2つ以上利用して、1つの理論変数を表現することもある

分析単位

変数を測定するにあたり、どこに注目するか
作業仮説で使われるすべての変数の分析単位は、同一でなければならない

Ex)収入が少ないほど共産党を支持する

① 分析単位:有権者個人
世帯収入→あなたの家族の年間所得は？
共産党支持→あなたは共産党に投票しましたか？

② 分析単位:都道府県
世帯収入→都道府県別平均所得
共産党支持→都道府県別共産党得票率

今回のケースでは、都道府県別のほうがコストは少なく済む

回帰分析のための仮説

・回帰分析:説明変数の値が応答変数(結果変数)の値に与える影響を調べる(説明変数の値に条件づけられた応答変数の期待値を推定するのが回帰分析)
・何が応答変数で何が説明変数かが明確にされた仮説が必要
・応答変数と説明変数が測定されていないといけない
・応答変数と説明変数が測定するものは別のものでなければいけない
・Rで分析するために:応答変数と説明変数を、それぞれ tidy dataの列として用意すべき

仮説検証のためのデータ

測定方法

実験が最善

世論調査

調査対象者に質問し、質問に答えてもらう
属性に関する質問:性別、職業、etc.
意見や態度に関する質問:Aに賛成？反対？
例:「イートインとテイクアウトで消費税率を変える こと(10%vs8%)に賛成ですか、反対ですか。」
知識:Bを知っているか？
例:「アメリカ合衆国、日本、英国、ドイツ、フランスを人口が多い順に並べ替えてください」

問題点

回答拒否:調査そのものを拒否する ＝ 全項目無回答(unit nonresponse)
調査を受け入れる人たちと拒否する人たちに違いがあると、サンプルの代表性が損なわれる

一部の質問に答えない ＝ 一部項目無回答 (item nonresponse)
他人に知られたくない情報は隠しがち(例:所得, イデオロギー)

難しい質問には答えない(わからない; DK [don't know])

真面目に答えない:satisficer(満足者)の問題

一部の回答者はウソをつく

・社会的な規範に反しない答えを選ぶ:**社会的望ましさバイアス(social desirability bias; SDB)**

・答えにくい質問でウソをつく

例:「前回の選挙で投票しましたか」という質問

✦8 割から 9 割が「投票した」と答える

✦実際の投票率は6割程度

SDB が疑われるような状況では、単純に質問しても欲しい答えが得られない(測定誤差)

→**アイテムカウント法(リスト実験)**

答えにくい質問に答えさせるための工夫

複数の項目を提示し、**該当する項目の数を答えてもらう**

回答者を実験群と統制群にランダムに割り当て、**実験群にだけ答えにくい質問項目を入れる**

統制群と実験群はランダムに割り当てているので、2つのグループは似ている(ほぼ同じ)はず。違いは、答えにくい質問項目が入っているかどうかだけ。**2つのグループの間で答えの平均数に違いがあれば、それは答えにくい質問の効果**のはず。つまり、**答えにくい質問に「はい」と答える人の割合が推定**できる。

世論調査の工夫

コンピュータによる回答で調査者に会わないようにする

質問の順番をランダム化する、尋ね方もランダムに変える

特定の回答者の票に傾斜配点つけて代表サンプルに近づける

誰を調査するか

研究の目的による

目的によって母集団は異なる

実際の対象者は、「標本抽出枠(sampling frame)」と呼ばれるリストから選ばれる

標本抽出枠が母集団と「ほぼ同じ」になるような工夫が必要

調査対象が…

母集団が小さい:全数調査

母集団が大きい:サンプル調査

ほとんどの世論調査がサンプル調査

標本の選び方

母集団の偏りのない縮図が欲しい　どうやってサンプリングするか

意図的に選ぶと、バイアスがかかってしまう

単純無作為抽出

母集団から標本をランダムに選ぶこと

無作為抽出で選び出された標本は、母集団の偏りのない縮図であるとみなすことができる。ただし、誤差はつきもの。

標本の数≠標本サイズ

標本の数：母集団から取り出した集団の数

標本サイズ（N）：1つの標本に含まれる個体の数

誤差

標本から得られる統計量は必ず誤差がある。

問題は、その誤差に偏りがない（誤差の平均値がマイナス以下）か、その誤差が小さいかが重要

回帰分析における統計的推測Ⅱ　～仮説検証～

母集団の回帰直線と標本の回帰直線

データから作った散布図への直線(平面)の当てはめは、標本データの要約
しかし、興味があるのは母集団の特徴
どのような方法で、標本から母集団を推定する？

統計モデルを作る

自分が観察しているデータが生み出される過程をモデル化する(データ生成過程)
モデル:目的に応じた事象の単純化　→単純化するので正確ではないが理解に役立つ

単回帰

母集団における単回帰

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

α, β:パラメタ，母数(推定の対象)
ε:誤差

誤差のモデル化

$$\varepsilon_i \sim Normal(0, \sigma)$$

誤差の平均は0
誤差は、1つの正規分布から生み出される
誤差の標準偏差σは、iによらず一定

単回帰モデル

単回帰が想定するDGP
まず$X_i (i = 1,2, \dots)$が決まる。このあとに$Y_i (i = 1,2, \dots)$が以下のように決まる

$$Y_i \sim Normal(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta X_i$$

別の書き方をすると

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$
$$\varepsilon_i \sim Normal(0, \sigma)$$

最小二乗法による母集団の単回帰分析

標本データを使い、最小二乗法よって求めた回帰係数a,bは、
単回帰モデルに登場する α, β の点推定値
最小二乗推定量は以下の望ましい性質をもつ

不偏性

一致性：標本サイズを無限大にすると、推定値は母数に到達する

最小二乗法による母集団の重回帰分析

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

$$\varepsilon_i \sim Normal(0, \sigma)$$

β_k：パラメタ，母数（推定の対象）

k=0, 1, 2, …, K

最小二乗推定量は以下の望ましい性質をもつ

不偏性 $\mathbb{E}[b_k] = \beta_k (k = 0, 1, 2, \ldots, K)$

一致性：標本サイズを無限大にすると、推定値は母数に到達する

回帰分析の帰無仮説と対立仮説

帰無仮説と対立仮説

帰無仮説：「説明変数は応答変数に影響を与えない」

対立仮説：「説明変数が応答変数に影響する」

自分が「正しい」ことを示したい理論の**作業仮説を対立仮説にする**

統計的検定（方法は後で説明する）で**帰無仮説が棄却されたとき、「作業仮説が統計的に正しい」と判断する**

つまり$\beta = 0$

重回帰の場合［包括的検定］

帰無仮説：

対立仮説：少なくとも一つにおいて$\beta \neq 0$

重回帰の場合［個別的検定］

モデル：$Y_i \sim \text{Normal}(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_K X_{iK}, \sigma)$

検証する仮説のパタン2

	β_1の仮説	β_2の仮説	…	β_Kの仮説
帰無仮説：	$\beta_1 = 0$	$\beta_2 = 0$	…	$\beta_K = 0$
対立仮説：	$\beta_1 \neq 0$	$\beta_2 \neq 0$		$\beta_K \neq 0$

- 実際は、すべての k について仮説を立てて検証するわけではなく、理論における「原因」とみなされるものについてのみ個別に仮説を検証する

影響力がない・効果がない は検証不可

通常、**「影響がない」は帰無仮説**

「影響がない」を対立仮説にすると、帰無仮説「影響がある」 は棄却できない（検証する対象が無限にある）

「影響がない」という帰無仮説を棄却できなくても、それは 「影響がない」ことを意味しない

「影響がある」という証拠が見つからないだけ

「証拠の不在」は「不在の証拠」ではない！

「影響がない」ことを主張する理論は、（これまで勉強してきた）統計的分析では検証不可

回帰分析における仮説検定

回帰分析では、説明変数が応答変数に影響を与えているかどうかに関心がある

帰無仮説：説明変数の影響はない（影響が0である）

対立仮説：説明変数の影響がある（影響が0ではない）

推定値のばらつき（単回帰の場合）

b のばらつき：

標本ごとに異なるβの標準偏差：標準誤差（SE）

$$\mathrm{SE}(b) = \sqrt{\frac{\hat{V}_1}{N}}$$

$$\hat{V}_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}\left[(x_i - \bar{x})^2 e_i^2\right]}{\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^2}$$

ただし、e_i は残差：$e_i = y_i - \hat{y}_i = y_i - (a + bx_i)$

推定量βの分布

$$\frac{b-\tilde{\beta}}{\mathrm{SE}(b)} \sim t(N-K-1)$$

- $\tilde{\beta}$：帰無仮説が想定する β
  - 帰無仮説が正しいなら、$\mathbb{E}[b]=\tilde{\beta}$
- $t(N-K-1)$：自由度 $N-K-1$ の t 分布
  - N：標本サイズ
  - K：説明変数の数（切片は含まない）

t 統計量を用いた仮説検定

t 統計量：$T=\dfrac{b-\tilde{\beta}}{\mathrm{SE}(b)}$

特定の有意水準のもとで、自由度 $N-K-1$ のt 分布の臨界値 c を求め、

$$|T|>|c|$$

となるとき、帰無仮説を棄却する

帰無仮説が $\beta=0$（つまり、$\tilde{\beta}=0$）のとき、

$$T=\frac{b-\tilde{\beta}}{\mathrm{SE}(b)}=\frac{b}{\mathrm{SE}(b)}$$

この T の値は、Rで回帰分析結果に t value または statistic として表示される

有意水準が5パーセントのとき、検定の臨界値は約2

- よって、係数を標準誤差で割った値の絶対値が2より大きければ、有意水準5%で帰無仮説を棄却する

信頼区間

回帰分析による点推定値は、1つの標本（データ）から得られたもの
→母数に一致するとは限らない（実際の標本サイズは有限なので）
統計量はばらつく（シミュレーションで確認する！）
標準誤差：統計量のばらつき ➡ 信頼区間を求める

95%信頼区間

1. データを生成する（新たに観測する）
2. データを分析する

3. 95%信頼区間を求める

上の1～3までを何度も何度も繰り返し行うと、そのうち95%は「真の値を含む信頼区間」が得られるだろう

＝95%の信頼区間に母数が入る

信頼区間の特徴

信頼度が高いほど区間が長くなる：区間を長くすれば、取りこぼしの確率が小さくなる

信頼度が低いほど区間が短くなる：区間を短くすれば、取りこぼしの確率は大きくなる

ただし、同じ信頼度で、信頼区間が短いほうが推定の不確実性が小さい

信頼区間の長さは標準誤差に依存するので…

標準誤差が大きい：信頼区間が長い

標準誤差が小さい：信頼区間が短い

統計的に有意とは

統計的に有意 ＝ **効果が 0 ではない** というだけ。

0ではないということが、研究においては重要視される

実質的重要性（＝どれくらい効果が大きいか小さいか）は「統計的に有意か」には関係ない

→すなわち係数の値そのものの議論が必要

「統計的に有意である」という結論は、ほぼ無意味

そこから実質的重要性などを掘り下げ、リサーチクエスチョンに答える必要

→「統計的・実質的に有意」「統計的に有意だが、実質的に無意味」「統計的に有意ではなかった」の3択の答えになるはず

因果推論

因果推論とは何か

学問の目的は、原因の結果(＝特定の原因によってどのような結果が生じる)を考えること

共変関係

　変数xが変化すると変数yが変化する

　相関関係も因果関係もどちらも共変関係として扱う

偶然だったり、交絡因子が挟まっているものは因果関係とは言えない

潜在的結果アプローチ

Ex)アスピリンを飲んだら、頭痛が収まった

もしあのとき、飲まなかったら…？　→アスピリンを飲まなかったときは、頭痛は残っていたはず。

潜在的結果（potential outcomes）

　アスピリンを飲んだ場合の頭痛の状態

　アスピリンを飲まない場合の頭痛の状態

因果関係があるなら、潜在的結果が行動(処置、介入、操作)によって変わるはず

個体単位での潜在的結果:

頭痛のある個人がアスピリンを飲んだら、1時間後に頭痛は消えるか？

・個人　$i = \in \{1,2,\dots,N\}$

・処置(原因)　$D_i \in \{0,1\}$:飲まない　＝ 0，飲む　＝ 1

・結果　$Y_i \in \{0,1\}$:頭痛なし＝ 0，頭痛あり　＝ 1

$Y_i(D_i)$：処置が D_i の場合の潜在的結果

$$Y_i = Y_i(1) \quad \text{if} \quad D_i = 1$$

$$Y_i = Y_i(0) \quad \text{if} \quad D_i = 0$$

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

$$= Y_i(0) + D_i[Y_i(1) - Y_i(0)]$$

1. アスピリンを飲んだ場合のみ頭痛が消える（薬の効果を示す**因果関係**）

$$Y_i(1) = 0,\ \ Y_i(0) = 1$$

2. いずれにせよ頭痛は残る

$$Y_i(1) = 1,\ \ Y_i(0) = 1$$

3. いずれにせよ頭痛は消える

$$Y_i(1) = 0,\ \ Y_i(0) = 0$$

4. アスピリンを飲んだ場合のみ頭痛が残る（逆の因果関係）

$$Y_i(1) = 1,\ \ Y_i(0) = 0$$

つまり、パターン1が正しいかを確かめたい。

因果効果は潜在的結果の差

ただし、同一個体の同一日時でのみ因果効果を認める。

ダメな因果推論

処置前と処置後を比較する	上記パターン3があり得てしまうから
異なる個体を比較する	違う環境にある

分析単位

処置（行動）は、分析単位（unit）に適用される

分析単位は

- 物理的対象：人、物
- 行政単位：国、県、市町村、州
- 物や人の集合（グループ）など

分析単位は、**「特定の時間」において定義され、同一人物でも、異なる時点では異なる単位として扱われる**

因果推論の根本問題とその克服

潜在的結果は、**タイムマシンでもない限り観察不能**なものである

→**個体の因果関係は絶対に観察不能**

克服のため複数の個体を考える

平均因果効果(平均処置効果)(ATE)

処置の値が 2 種類(0 か 1)しかないとき

処置 1 を受ける:処置を受ける

　処置を受けた個体のグループ:処置群（treatment group)、実験群

処置 0 を受ける:処置を受けない

　処置を受けなかった個体のグループ:統制群

にわけて調べる。

ただし、処置 1 を受けた個体と処置 0 を受けた個体がいるとき、どちらの期待値も観察(推定)できない　→**平均因果効果は観察できない**

処置群における平均因果効果(ATT)

ATE がダメなので、観察された平均値を比較しよう

式変形したら…

$$\mathbb{E}[Y(1) \mid D=1] - \mathbb{E}[Y(0) \mid D=0]$$

$$= \mathbb{E}[Y(1) \mid D=1] - \mathbb{E}[Y(0) \mid D=0]$$

$$+ (\mathbb{E}[Y(0) \mid D=1] - \mathbb{E}[Y(0) \mid D=1]) \quad \leftarrow 0$$

$$= \mathbb{E}[Y(1) \mid D=1] - \mathbb{E}[Y(0) \mid D=1] \quad \leftarrow \text{ATT}$$

$$+ \mathbb{E}[Y(0) \mid D=1] - \mathbb{E}[Y(0) \mid D=0]$$

$\leftarrow$ セレクションバイアス

セレクションバイアス:個体ごとの違いによって発生するバイアス(内生性)

処置を受ける(処置 1)か、処置を受けない(処置 0)かが、結果の値によって異なる

Ex)手術がうまくいきそうな人ほど手術を受け、手術が失敗しそうな人ほど手術を避ける

- $\mathbb{E}[Y(0) \mid D=1]$：処置を受けた群の個体が、処置を受けなかったときの潜在的結果の期待値
- $\mathbb{E}[Y(0) \mid D=0]$：処置を受けなかった群の個体が、処置を受けなかったときの潜在的結果の期待値

- $\mathbb{E}[Y(0) \mid D=1] = \mathbb{E}[Y(0) \mid D=0]$　ならセレクションバイアスはない

  → その場合、ATT が推定できる (ATE ではないので注意)

- バイアスがある：処置の値と潜在的結果の値に相関がある
  - 処置を受けた群と受けていない群で、結果のベースラインに違いがある

セレクションバイアスに対処する（なるべく0に近づける）

無作為化比較試験（RCTs）

個体を処置群と統制群に無作為に振り分ける（＝処置変数 D を完全ランダム化）

$$\mathbb{E}[Y(0) \mid D=1] = \mathbb{E}[Y(0) \mid D=0]$$

かつ

$$\mathbb{E}[Y(1) \mid D=1] = \mathbb{E}[Y(1) \mid D=0]$$

したがって、

$$\begin{aligned}&\mathbb{E}[Y(1) \mid D=1] - \mathbb{E}[Y(0) \mid D=0]\\ &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]\\ &= \mathrm{ATE}\end{aligned}$$

$$\mathbb{E}[Y(1)] = \mathbb{E}[Y(1) \mid D=1] = \mathbb{E}[Y(1) \mid D=0]$$

かつ

$$\mathbb{E}[Y(0)] = \mathbb{E}[Y(0) \mid D=1] = \mathbb{E}[Y(0) \mid D=0]$$

観察したものから、ATE が推定できる！

無作為に作られる2つの集団（処置群と統制群）は、よく似ているので交換可能とされる

処置群と統制群の違いを処置行為のみにすることで、比較ができる。

実験ができないとき

調査・観察データを使った因果推論は難しい

処置を受ける人と受けない人が「同じ」ではないから実験ができない

交絡

第三の要因のこと

Ex）性別　男性だから運動たくさんするので、体の消費も激しく、寿命は相対的に短くなる？

交絡の防止のために

ブロッキング（≒重回帰分析）、再分類化

交絡変数の値によって、分析対象をグループ分けして分析する

性別が交絡なら、男性と女性を別に分析する

回帰分析の応用

統計的因果推論

因果関係がはっきりとさせるには実験がベスト

倫理的、コスト的な問題や過去のデータを調べるときなどには実験などはできない：準実験

- **自然実験**（natural experiments）
- **準実験**（quasi-experiments）
  - **操作変数法**（instrumental variable method）
  - **回帰不連続デザイン**（regression discontinuity design）
  - **差分の差分法**（difference-in-differences [DiD]）
- **条件付け**
  - **統制変数を伴う回帰分析**
  - **パネルデータ分析**

社会科学において、調べることの多くは確率的傾向（＝傾向にある）

そのため、大きな標本サイズで、処置群と統制群をそれぞれ平均して比較することが必要

重回帰分析

「他の条件が等しい」状況を作り出すため、重回帰分析を利用する

検証したい理論：「X が Y を上昇（減少）させる」

応答変数：Y

主な説明変数:X

統制（コントロール）変数:Z（複数あってよい）

私たちが比較したい個体が様々な面で異質なとき、複数の Z 要因を統制する必要がある

理論的関心：X が Y に影響するかどうか

重回帰モデル：$Y_i \sim \text{Normal}(\alpha + \beta X_i + \gamma Z_i, \sigma)$

β の推定値：Z の影響を取り除いたとき、X 1単位の増加がY を何単位増加させるか

γ の推定値：意味なし！

- Z が**コントロール変数**なら、γ の**意味を解釈しようとしてはいけない！**

理論的関心：Y に影響を与える変数は何か？

重回帰モデル：$Y_i \sim \text{Normal}(\alpha + \beta X_i + \gamma Z_i, \sigma)$

β の推定値：Z の影響を取り除いたとき、X 1単位の増加がY を何単位増加させるか

γ の推定値：X の影響を取り除いたとき、Z 1単位の増加がY を何単位増加させるか

ガンマの解釈の有無は、問いによって変わる

統制変数の選び方

重回帰分析には、入れるべき変数と入れてはいけない変数がある

- **統制すべき変数**

交絡変数

統制してはいけない変数

処置後変数：処置した後に変化する変数

この処置後変数は重回帰分析から取り除く必要があり、これを入れると処置後変数バイアスが発生する

媒介変数

合流点

交絡の仕方

Y が結果、X が原因のとき

1. Z は X と Y の交絡因子である
2. Z は X と Y の結果として生じる（＝**合流点**）
3. Z は X と Y の**媒介変数（mediator，中間因子）**である（X が Y にも影響を与えるが、同時に Z 経由でも Y に与える）

重回帰分析には Z を固定することで、比較ができるようにする。

最小二乗法で推定した線形回帰モデルの場合

Y：応答変数

X:主な説明変数

Z:統制変数（共変量）

私たちが知りたい（推定する）のは、X が Y に与える影響

X の Y に対する因果効果：X が 1 単位増加したとき、 Y は何単位増加するか？

この効果を推定する：係数の点推定値と信頼区間を出す

最小二乗推定量が、因果効果の推定を誤る場合

推定結果にバイアスが生じる

内生性（endogeneity）

（欠落［脱落］変数バイアスやセレクションバイアスともいう）

変数変換

回帰式をより解釈しやすいものにするために、変数を変換する

1 次関数を利用して変換する

測定単位の変更

Ex）100万円あたりを1円あたりに変更

標準化

すべての説明変数を Z 値で標準化する

$$z(x) = \frac{x - \bar{x}}{u_x} = \frac{x - x\text{の平均値}}{x\text{の不偏分散の平方根}}$$

回帰係数：他の説明変数の値を一定に保ち、注目する説明変数 X の値を 1 標準偏差分大きくしたとき、応答変数 Y が何単位分大きくなるか

切片が、「すべての説明変数がそれぞれの平均値をとったときの応答変数の予測値」という意味を持つようになる。

その他の標準化

単位を変える

- その他の例：ある意見に賛成か反対かを7点尺度で尋ねる
  - 1点：強い反対，・・・，7点：強い賛成：回帰係数の解釈が難しい
  - 標準化する $\frac{\text{得点} - 4}{3}$
    - -1点 = 強い反対, 0点 = 中立, 1点 = 強い賛成
    - 回帰係数：強い反対と中立の差、中立と強い賛成の差

スケーリングの方針

どの単位で測ることに意味があるか？

　Ex)一円変わったところで大きな変化が生じないものは、100万円単位に変えてみる

重回帰の場合：係数の値が変数ごとにあまりにも大きくばらつくことを避ける

中心化

0 に意味がない説明変数が存在

その場合、回帰式の切片に意味がなくなる。

中心化は、回帰式の切片に意味を持たせるためにやる

- 標本平均を使った中心化

$$x_c = x - \bar{x}$$

- 基礎知識や慣習を使った中心化
  - 例1）女性ダミーの中心化：男女比が1対1だと仮定

$$\text{female}_c = \text{female} - 0.5$$

  - 例2）知能指数 (IQ) の中心化：平均は100のはず

$$\text{IQ}_c = \text{IQ} - 100$$

- すべての説明変数が中心化された回帰式の切片：すべての説明変数が平均（またはその他の中心）の値をとったときの応答変数の予測値（平均値）

標準化した変数による単回帰

標準化された変数 z_x と z_y を用いた単回帰：

$$z_{yi} \sim \text{Normal}(s + rz_{xi}, \sigma)$$

ただし、$Z_{yi} = \dfrac{y_i - \bar{y}}{u_y}$, $Z_{xi} = \dfrac{x_i - \bar{x}}{u_x}$

切片 $s = 0$

傾き　$r \in [-1,1]$：x と yの相関係数

つまり、通常の回帰：$y_i = a + bx_i + e_i$ で

$$|b| > 1 \Rightarrow u_y > u_x$$

相関係数と単回帰の回帰係数の関係

- 一般的な単回帰（標準化されていない場合）を考える　傾きと相関係数が右記であることから一致する
  - x と y の共分散を $\text{Cov}(x, y)$ とする
  - x と yの相関係数 r：$r = \dfrac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)\text{Var}(y)}}$
  - 回帰式の傾き b：

$$b = \frac{\text{Cov}(x, y)}{\text{Var}(x)} = \frac{\text{Cov}(x, y)}{\text{Var}(x)} \frac{\sqrt{\text{Var}(x)}}{\sqrt{\text{Var}(x)}} \frac{\sqrt{\text{Var}(y)}}{\sqrt{\text{Var}(y)}}$$
$$= r\frac{\sqrt{\text{Var}(y)}}{\sqrt{\text{Var}(x)}} = r\frac{\text{SD}(y)}{\text{SD}(x)}$$

主成分線

線そのものと観測値との差が最も小さくなる線

x が小さいと、y が過小評価され、x が大きいと y が過大評価されるという欠陥がある

（⇔回帰直線：どの x の周辺でも、データの中心を予測）

- 平均への回帰：標準偏差で測ったとき、

$$\hat{y} \text{ と } \bar{y} \text{ の距離} < x \text{ と } \bar{x} \text{ の距離}$$

  - 「どんな変数も次第に平均に近づく」とは**言っていない**
  - 予測値の平均値からの乖離は、説明変数の平均値からの乖離より小さい（割り引いて考える）ということ

対数

指数関数の逆関数

対数変換することで、結果が安定し、わかりやすくなる場合がある

- x の自然対数：$\log_e(x) \rightarrow$ 単に $\log(x)$ と書く
  - 経済学では、$\ln x$ とされることも多い
- 自然対数を使う理由：結果がわかりやすい
- 例：応答変数が自然対数のとき

$$\log(y_i) = b_0 + 0.06x_i + e_i$$

  - x が1単位増えると、$\log(y)$ は0.06単位増える
  - x が1単位増えると、y は $\exp(0.06) - \exp(0) = \exp(0.06) - 1 \approx 0.06$ 単位増える
  - x の1単位分の増加は、y を約6%（つまり、0.06）増加させる
  - 係数 0.06：y の変化率（ただし、この近似が使えるのは、係数が0に近いときだけ）

応答変数が自然対数のとき

係数が 0 に近いときは、係数を変化率と考えることができる

$$\log_{10}(y_i) = b_0 + 0.026x_i + e_i$$

x が1単位増えると、$\log_{10}(y)$ は0.026単位増える

x が1単位増えると、y は

$10^{0.026} - 10^0 = 10^{0.026} - 1 = 0.06$ 単位だけ増える

x の1単位分の増加は、y を約6%（つまり、0.06）増加

させる

係数 0.026：このままでは、y の変化率はわからない！

変化率を調べたい場合、対数変換して回帰分析するとこのようにわかりやすくなる
経済学ではよく使われる

様々な仮説検定

二次式の推定

線形回帰で二次関数はできる？
→X と Z という2つの説明変数を持つ重回帰分析としてあらわすことができる
ただし、β_1の推定値は、Z の値が一定になったときに X が Y に与える影響とは解釈できないことに注意

分析結果の提示法

回帰分析の結果の提示

図、表または式の形で表す

係数だけでなく、不確実性(**標準誤差**, 値, t 値)も一緒に示すことが必要

どの不確実性指標を使っているかはっきり示すこと！

標準誤差を示すのがもっとも望ましい

点推定値と信頼区間を図示する

観測数(サンプルサイズ)と決定係数(重回帰の場合は自由度調整済み決定係数)も示す

決定係数 R^2

Summary 関数を出すと…

```
Call:
lm(formula = voteshare ~ experience, data = HR09)

Residuals:
    Min      1Q  Median      3Q     Max
-43.867 -12.072  -5.567   8.583  52.123

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.8772     0.6203   22.37   <2e-16
experience   30.9898     0.9783   31.68   <2e-16

Residual standard error: 16.19 on 1137 degrees of freedom
Multiple R-squared:  0.4688,    Adjusted R-squared:  0.4684
F-statistic:  1004 on 1 and 1137 DF,  p-value: < 2.2e-16
```

Multiple R-squared: 0.4688 が R 決定係数

Adjusted R-squared: 0.4684 は自由度調整済み R^2 係数　重回帰分析の際には提示する必要がある

式の書き方

$$\widehat{\text{身長}} = 107.2 + 0.19 \times \text{父の身長} + 0.21 \times \text{母の身長}$$
$$\qquad\qquad (4.93) \quad (0.02) \qquad\qquad\qquad (0.02)$$

注：括弧内は標準誤差

括弧内には、**標準誤差（se）を書く**

標準誤差が書かれている場合の目安：有意水準 5%なら、係数 ÷SE の値が 2 以上なら帰無仮説（=0）を棄却

t 値（検定統計量）を書いても理論的には問題ないが、標準誤差のほうが信頼区間を計算しやすい

図示することもできる。（付録にして書いたほうが良き）

重回帰分析の場合の書き方の注意

複数ある説明変数のうち、注目する変数は限られている

交絡変数の推定値の意味は解釈できないので、報告しない

ただし、表を付録に載せる場合は、交絡についての推定値も載せておく

注目する説明変数が 2 つ以上ある場合は、それぞれについて丁寧に説明する

交差項がある場合は要注意（後述）

推定値をそのまま報告するだけではダメ

交差項の処理

交差項を含む回帰分析

説明変数が応答変数に与える影響は一定ではない場合がある

Z が調整変数の場合

Z によって Y への影響が変わる

- 回帰モデル：$Y_i \sim \text{Normal}(\mu_i = \beta_0 + \beta_1 X_i, \sigma)$
  - β_0 は Z の関数：$\beta_0 = \gamma_0 + \gamma_2 Z_i$ とする
  - β_1は Z の関数：$\beta_1 = \gamma_1 + \gamma_3 Z_i$ とする

$$\begin{aligned}\mu_i &= \beta_0 + \beta_1 X_i \\ &= (\gamma_0 + \gamma_2 Z_i) + (\gamma_1 + \gamma_3 Z_i) X_i \\ &= \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \gamma_3 X_i Z_i\end{aligned}$$

  - よって、回帰モデルは、

$$Y_i \sim \text{Normal}(\gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \gamma_3 X_i Z_i, \sigma)$$ ←ここで重回帰分析にできる

- $Y_i \sim \text{Normal}(\gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \gamma_3 X_i Z_i, \sigma)$
  - Y を X と Z と XZ に回帰する重回帰
    - γ_k $(k = 0,1,2,3)$ を推定し、**そこから β_1 を推定する**
  - $\gamma_3 X_i Z_i$：交差項, 交互作用項 (interaction term)
    - Rでは、`lm(Y ~ X * Z, data = d)`

Z がダミー変数の場合

$$Y_i \sim \text{Normal}(\mu_i = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \gamma_3 X_i Z_i, \sigma)$$

- $Z_i = 0$ のとき：

$$\hat{Y}_i = \gamma_0 + \gamma_1 X_i + \gamma_2 \cdot 0 + \gamma_3 \cdot X_i \cdot 0 = \gamma_0 + \gamma_1 X_i$$

- $Z_i = 1$ のとき：

$$\hat{Y}_i = \gamma_0 + \gamma_1 X_i + \gamma_2 \cdot 1 + \gamma_3 \cdot X_i \cdot 1 = \gamma_0 + \gamma_1 X_i + \gamma_2 + \gamma_3 X_i$$
$$= (\gamma_0 + \gamma_2) + (\gamma_1 + \gamma_3) X_i$$

- Z の値によって、
  - 回帰直線の「切片」が変わる（$\gamma_2 \neq 0$ のとき）：γ_0 or $\gamma_0 + \gamma_2$
  - 回帰直線の「傾き」が変わる（$\gamma_3 \neq 0$ のとき）：γ_1 or $\gamma_1 + \gamma_3$

$$Y_i \sim \text{Normal}(\mu_i = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \gamma_3 X_i Z_i, \sigma)$$

- $Z_i = 0$ のとき：切片は γ_0、傾き（X が Y に与える影響）は γ_1
- $Z_i = 1$ のとき：切片は $\gamma_0 + \gamma_2$、傾き（X が Y に与える影響）は $\gamma_1 + \gamma_3$

★ 推定された偏回帰係数の意味

- γ_0：$Z = 0$ のときの回帰直線の切片
- γ_1：$Z = 0$ のときの回帰直線の傾き
- γ_2：$Z = 1$ のときと $Z = 0$ のときとの回帰直線の切片の差
- γ_3：$Z = 1$ のときと $Z = 0$ のときとの回帰直線の傾きの差

Z が量的変数のとき

$$Y_i \sim \text{Normal}(\mu_i = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \gamma_3 X_i Z_i, \sigma)$$

- X-Y 平面における回帰直線の切片：$\gamma_0 + \gamma_2 Z_i$
- X-Y 平面における回帰直線の傾き：$\gamma_1 + \gamma_3 Z_i$

★ 切片も傾きも Z_i の値によって変わる！

- $Z_i = 0$ のとき：切片は γ_0、傾き（X が Y に与える影響）は γ_1
- $Z_i \neq 0$ のとき：切片も傾き（X が Y に与える影響）も、Z_i の値による
  - **回帰係数だけを見ても、X が Y に与える影響はわからない！！！**
- **Z_i を横軸、$\gamma_1 + \gamma_3 Z_i$（X が Y に与える影響）を縦軸にした図を作る！(A)**
- **Z_i をいくつかの値に固定して、複数の回帰直線を図示する！ (B)**

すべてzに依存することになるので、統計的優位性は語れない
（以下に続く）

$$Y_i \sim \text{Normal}(\mu_i = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \gamma_3 X_i Z_i, \sigma)$$

★ 推定された偏回帰係数の意味

- γ_0：Z＝0 のときの回帰直線の切片
- γ_1：Z＝0 のときの回帰直線の傾き
- γ_2：？（説明しようと思えば説明できるが、わかりにくい）
- γ_3：？（説明しようと思えば説明できるが、わかりにくい）

・Z_i が0をとらない変数だったら？？？

- それぞれの偏回帰係数に意味がない：回帰係数を表で提示しても、読者に意味が伝わりにくい

交差項を出すなら、中心化しておいたほうが見やすくなる（好み）

$$Y_i \sim \text{Normal}(\tilde{\mu}_i = \tilde{\gamma}_0 + \tilde{\gamma}_1 \tilde{X}_i + \tilde{\gamma}_2 \tilde{Z}_i + \tilde{\gamma}_3 \tilde{X}_i \tilde{Z}_i, \sigma)$$

- $\tilde{X}_i$：X_i を中心化したもの
- $\tilde{Z}_i$：Z_i を中心化したもの

★推定された偏回帰係数の意味

- $\tilde{\gamma}_0$：$\tilde{Z}=0$ のとき、すなわち Z が平均値のときの回帰直線の切片
- $\tilde{\gamma}_1$：$\tilde{Z}=0$ のとき、すなわち Z が平均値のときの回帰直線の傾き
- $\tilde{\gamma}_2$：？（説明しようと思えば説明できるが、わかりにくい）
- $\tilde{\gamma}_3$：？（説明しようと思えば説明できるが、わかりにくい）

・中心化することによって、$\tilde{\gamma}_0$ と $\tilde{\gamma}_1$ の意味だけは解釈可能になることが保証される

交差項の重要性

交差項を除くことは、γ3=0が証明できない限り除けないので基本的に入る。（同時に重回帰分析化することになる）

交差項を使うときは、交差項を構成するそれぞれの変数も説明変数に加える（理論的に正当化できる場合は除く）
交差項があるときは、偏回帰係数だけでは意味がわからない（結果を表で示すだけで不十分！）
効果を可視化（作図）する！