高知工科大学 経済・マネジメント学群

計量経済学応用

8. 回帰不連続デザイン

矢内 勇生（やない ゆうき）

https://yukiyanai.github.io

yanai.yuki@kochi-tech.ac.jp

このトピックの目標

- 回帰不連続デザイン (RDD) の考え方を理解する
  - RDDで因果推論できるのはなぜ？
  - RDDで必ず使われる結果の可視化法とは？
  - RDDに必要な仮定は？
- RDDによって因果効果を推定する方法を理解する
  - ノンパラメトリックな方法
  - パラメトリックな方法

回帰不連続デザイン

回帰不連続デザイン

- 回帰不連続デザイン、非連続回帰デザイン、regression discontinuity design (RDD)、RDデザイン
- 強制的な（自然法則ではない）ルールによって生まれる境界線を利用する
  - 法定飲酒年齢
  - 小学校での1クラスの人数の上限
  - 選挙区の定数
- 境界線を生み出すルールを、ランダム割付けに見立てる
  - 処置群：ルールによって境界線を「越える」個体の集合
  - 統制群：ルールによって境界線を「越えない」個体の集合

★ 自然実験（natural experiment）、準実験 (quasi-experiment)

境界線に注目する (1)

- 境界線付近の個体は「ほぼ」同じ
- 例：飲酒が死亡率に与える影響
  - 法定飲酒年齢が20歳だとする
    - 20歳になる1日前の人：飲酒できない
    - その日に20歳になった人：飲酒できる
  - 飲酒できるかどうか以外に、「死」に影響を与える違いはないはず

★ 20歳の誕生日という境界線の前後で「死亡率」に違いがあれば、それは「飲酒」の効果だとみなせるのでは？

境界線に注目する (2)

- 結果変数：テストの点数
- 処置：少人数クラス
- 割当変数：1学年の人数
  - 1クラスの上限人数が40だとする
  - 1学年の人数が40人だと40人の1クラス；41人だと20人と21人の2クラス
  - 1学年の人数が80人だと40人ずつの2クラス；81人だと27人ずつの3クラス

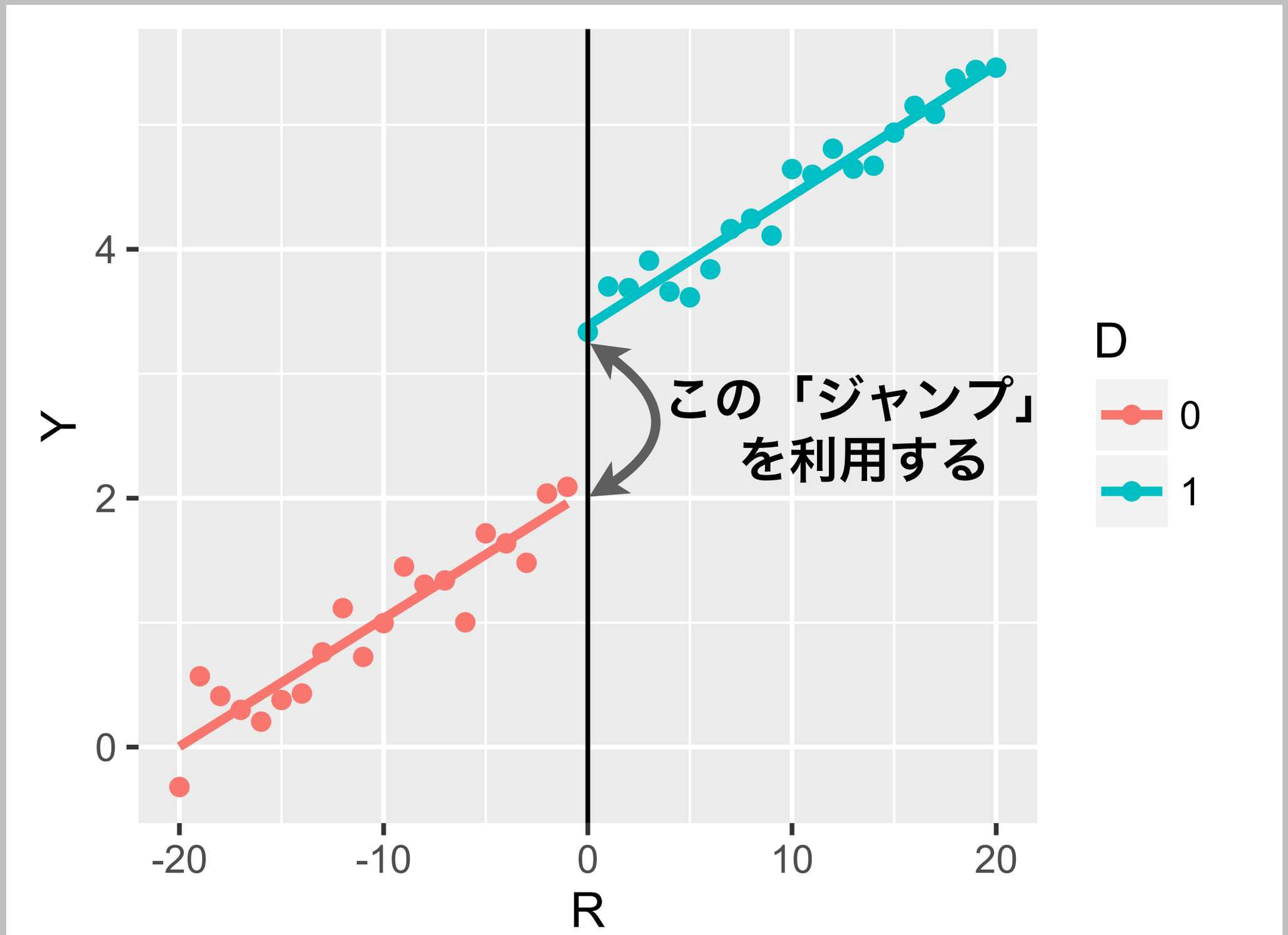
    etc.
- このような例をたくさん集めれば、40人クラスと20 [21] 人クラスには、平均すればクラスサイズ以外にテストの点数に影響を与える違いがないはず

★ テストの点数に違いがあれば、それはクラスサイズの影響では？

RDD で用いる変数

- 結果変数：Y_i
- 処置変数：D_i
- 割当変数 (running variable; forcing variable)：R_i
  - R は D と Y の両者に影響を与える
- 推定したい効果：D が Y に与える影響

例：線形の非連続回帰



例：非線形の非連続回帰

中室・津川 (2017: p.140) 図表6-2

RDDで推定する因果効果

- 「境界線付近」で処置が結果に与える効果
  - 境界線上の「ジャンプ」の大きさ
  - 「境界線付近の」局所的平均処置効果 (local average treatment effect; LATE)
    - 処置群と統制群が「ほぼ同じ」とみなせるのは、境界線に「近い」場合のみ
- 境界線を決める割当変数 R の値 R_c
  - カットオフ値 (cutoff, cut point) 、閾値(しきいち)

RDDの仮定 (1)

- 境界線によって処置の値が変わらない（境界線を作るルールが存在しない）場合、結果変数の値が非連続的な変化（ジャンプ）をすることはない
  - 例：法定飲酒年齢が20歳の場合
    - 疑問：20歳の誕生日に「ジャンプ」があるのは、誕生日パーティーで飲みすぎるせいでは？

✦ 法定飲酒年齢が20歳でなければ、20歳の誕生日にジャンプはなくなると仮定したい

❖ 19歳の誕生日、21歳の誕生日、22歳の誕生日 … にはジャンプがないことを示せば、説得力が増す

★ $\mathbb{E}[Y_i(0) \mid R_i]$ と $\mathbb{E}[Y_i(1) \mid R_i]$ が R_c で連続な関数

RDDの仮定（2）

- 境界線で「非連続的な変化」をするのは注目している処置変数のみ
  - カットオフ値の周辺で、結果に影響を与える他の「処置」がないことが必要
  - この仮定が破られる例
    - 20歳の誕生日に政府から運転免許証が自動的に発行される（注意：事実ではない）
    - この場合、死亡率が上がるのは飲酒のせいか急に運転したせいか区別できない

RDDの仮定（3）

- 割当変数の値を自分で選ぶことができない
  - 公立学校の1学年の人数：学校が人数を選ぶことはできない
  - 年齢：その個人が自分で選ぶことはできない
- 境界線付近で、割当変数の分布に不自然な変化がないか確かめることが必要
  - この仮定が破られる例
    - 身分証の偽造が簡単にできる（あるいは、アルコールを買うときに身分証の確認がない）なら、酒を買うための「年齢」は自分で選べる

RDDの長所

- RDDの前提条件が成り立つとき、境界線付近に限れば RCT が行われたとみなせる
  - 自然実験 (natural experiment)、準実験 (quasi-experiment)
- 図によって結果を示すことができる
  - 境界線における「ジャンプ」≈ LATE
- 現実社会に存在するさまざまな境界線に応用可能

RDDの短所

- RDDの仮定が正しいことを証明できない
- 因果効果を推定できるのが、境界線付近のみである
  - サンプル全体についての因果効果はわからない
- 実際の推定の場面では、
  - バンド幅によって推定結果が変わってしまう
  - 正しい関数形がわからないとバイアスが生じてしまう

RDDの詳細

2種類のRDD

- Sharp RD：割当変数の値が分かれば、処置の値が「確定」する場合
- Fuzzy RD：割当変数の値によって処置される確率が決まり、その確率が境界線上でジャンプする場合
- この授業では Sharp RDD のみ説明する

Sharp RD の特徴

- D の値が割当変数 R の値によって**確定**する

$$D_i = \begin{cases} 1 & \text{if } R_i \geq R_c \\ 0 & \text{if } R_i < R_c \end{cases}$$

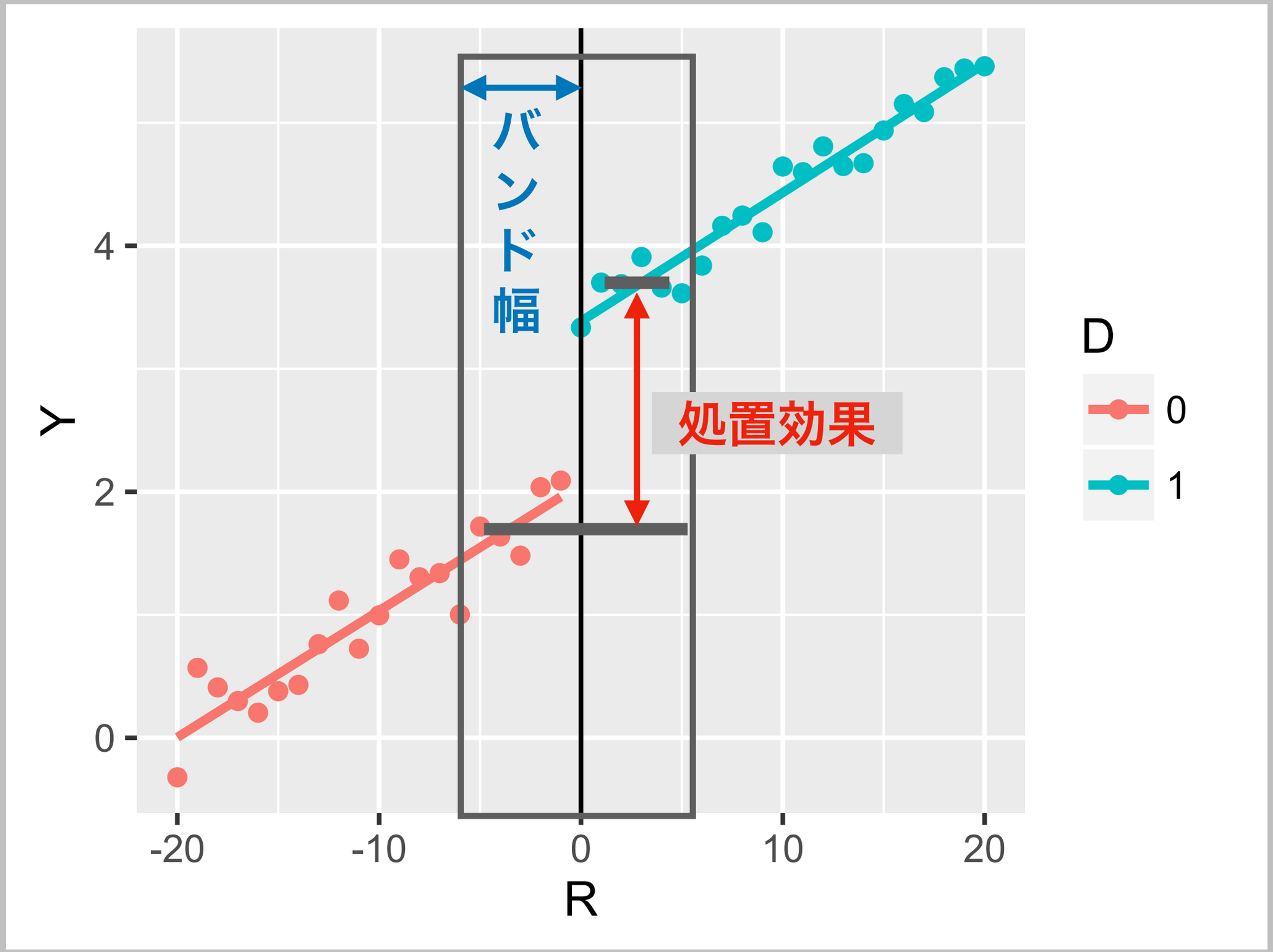
ここで、R_c はカットオフの値

- R の値のみで D が決まるので、R が唯一の交絡因子である
  - 欠落変数バイアスの心配がない
- R がどんな値であっても、その値で処置と統制の両者が観測されることはない (処置群と統制群の overlap がまったくない)
  - 処置群と統制群が比較できるのは、境界線上 [付近] だけ

RDDを用いた因果効果の推定

- 境界線上：処置群と統制群に処置以外の違いがない
  - 処置群と統制群で、結果変数の平均値を比較する
    - ノンパラメトリック (non-parametric) RD
- 境界線付近以外も含む推定
  - 回帰式の関数形を「正しく」特定することが必要
    - パラメトリック (parametric) RD

ノンパラメトリックRD



バンド幅 (bandwidth)

- バンド幅 (b)：ノンパラメトリックRD で推定に使うサンプルの範囲を決める幅
  - サンプルを $R_c - b < R_i < R_c + b$ となるものに限定し、処置群と統制群の平均値を比較する
- バンド幅を狭くすることで、「境界線付近」の比較を実行する
- バンド幅のトレードオフ：バンド幅を小さくするほど
  - 推定のバイアスが小さくなる
  - 標準誤差が大きくなる（サンプルサイズが小さくなる）
    - 詳しくは、Imbens & Kalyanaraman (2011) を参照

パラメトリックRD

- 特定の定式化による回帰式で処置効果を推定する
- 最も単純な回帰式（この関数形が正しいかどうかはわからない！）
  - $Y_i = \alpha + \rho D_i + \beta R_i + e_i$
    - R_i は交絡因子なのでコントロールすることが必要
    - 「この関数形が正しければ」ρ の推定値が処置効果の推定値

パラメトリックRDの潜在的結果

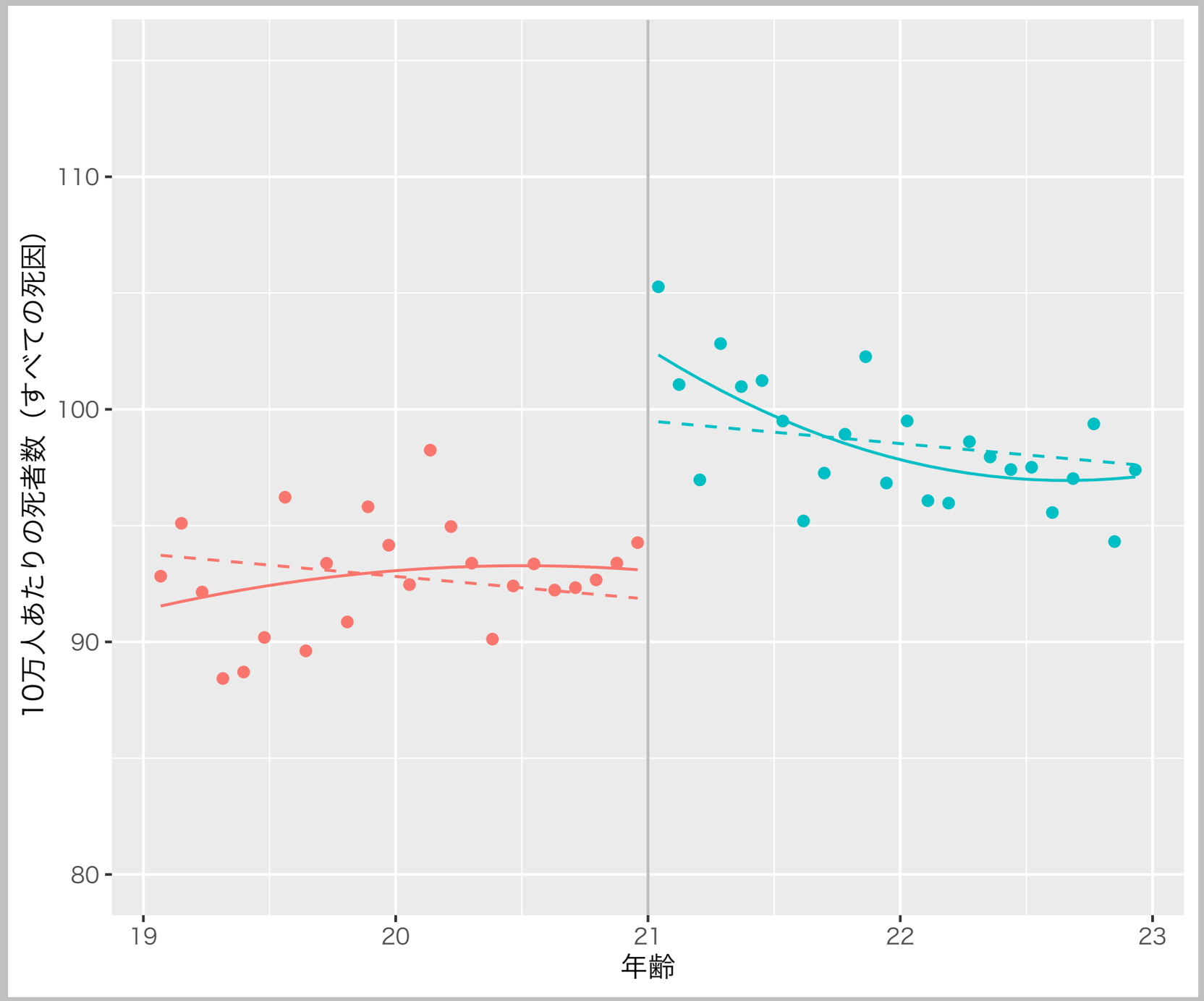
- 回帰関数：$\mathbb{E}[Y_i \mid D_i, R_i] = \alpha + \rho D_i + \beta R_i$
  - $D_i = 1$：$\mathbb{E}[Y_i(1) \mid D_i = 1, R_i] = \alpha + \rho + \beta R_i$
  - $D_i = 0$：$\mathbb{E}[Y_i(0) \mid D_i = 0, R_i] = \alpha + \beta R_i$
- 交絡は R のみ：条件付き交換可能性
  - $\rho = \mathbb{E}[Y_i(1) \mid R_i] - \mathbb{E}[Y_i(0) \mid R_i]$
  - $\mathbb{E}[\rho] = \mathbb{E}\left[\mathbb{E}[Y_i(1) \mid R_i] - \mathbb{E}[Y_i(0) \mid R_i]\right] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$
    - 問題：R に条件付けると、$Y_i(1) \mid D_i = 1$ または $Y_i(0) \mid D_i = 0$ のいずれか一方しか観察できない

推定結果は関数形に依存する

- パラメトリックRDでバイアスをなくすには、正しい関数形が必要
- 回帰式の例
  - $Y_i = \alpha + \rho D_i + \beta R_i + e_i$
  - $Y_i = \alpha + \rho D_i + \beta_1 R_i + \beta_2 R_i^2 + e_i$
  - $Y_i = \alpha + \rho D_i + \beta_1 R_i + \beta_2 R_i^2 + \beta_3 R_i^3 + \cdots + \beta_k R_i^k + e_i$
  - $Y_i = \alpha + \rho D_i + \beta R_i + \gamma (D_i \cdot R_i) + e_i$
  - $Y_i = \alpha + \rho D_i + \beta_1 R_i + \beta_2 R^2 + \gamma_1 (D_i \cdot R_i) + \gamma_2 (D_i \cdot R_i^2) + e_i$
  - $Y_i = \alpha + \rho D_i + \beta_1 R_i + \cdots \beta_k R^k + \gamma_1 (D_i \cdot R_i) + \cdots + \gamma_k (D_i \cdot R_i^k) + e_i$

  etc.

関数形の違いによる推定の差



Angrist & Pischke (2015: p.158) Figure 4.4 の再現

曲線の「フィット」を改善したいわけではない！

- 多項式を利用すれば、回帰曲線とデータの「フィット」はいくらでもよくできる
  - 標本サイズが2なら1次（直線）で完全にフィット
  - 標本サイズが3なら2次の多項式で完全にフィット
  - ・・・
  - 標本サイズが n なら n - 1 次の多項式で完全にフィット
- しかし高次の多項式をRDD に使うべきではない (Gelman & Imbens 2019)

3次の多項式の例

Chen et al. (2013) Fig. 3

非連続と非線形を区別せよ

Angrist & Pischke (2015: p.154) Figure 4.3

RDDのまとめ

- RDD：境界線を利用して因果効果を推定する
  - 最も重要な仮定：「境界線を作るルールがなければ、潜在的結果は連続」
  - 境界線付近のサンプルについての因果効果しかわからない
- 処置効果 (LATE) の推定法
  - ノンパラメトリックRD：バンド幅を決めて、境界線付近のサンプル内で処置群と統制群を比較
  - パラメトリックRD：全サンプルを使い、「正しい」関数形で処置効果を推定
- 「ジャンプ」を示す図が重要

次回予告

9. 操作変数法