



高知工科大学 経済・マネジメント学群

計量経済学応用

4. 回帰分析

やない ゆう き
矢内 勇生

🌐 <https://yukiyanai.github.io>

✉ yanai.yuki@kochi-tech.ac.jp



RCTの問題点

- どんな処置でもランダム化していいのか？
 - ▶ 病院に行くかどうか、実験者がコイントスで決めていいのか？
 - ▶ どんな処置を与えてもいいのか？
- ランダム化できない問題もあるのでは？
 - ▶ RCT ができない問題は研究できない・すべきでないのか？
 - ▶ 実験外の観察からしか得られない情報（データ）もあるのでは？
- ランダム化されていない処置の効果を推定したい！

今日の目標

- 因果推論に回帰分析を利用する方法を身につけよう
 - ▶ 回帰係数は条件付き期待値の差
 - ▶ 重回帰でセレクションバイアスを除去する
 - ▶ 回帰分析の「誤用」によるバイアス
 - 脱落変数バイアス
 - 処置後変数バイアス
 - ▶ DAG とバックドア基準

回帰分析について、今回説明しないこと

- 回帰分析の基本的な説明は「計量経済学」で学習済み
 - ▶ 因果推論についても少し説明したが、その部分は後で詳しく復習する
- 以下の内容は（おおむね）理解していると仮定する
 - ▶ 回帰分析とは何か
 - 回帰係数の求め方、最小二乗法
 - ▶ 回帰分析における統計的検定
 - 回帰分析で検証する仮説
 - 仮説の検証方法： p 値とは？
 - ▶ Rで回帰分析を実行する方法
 - `lm()` で回帰式を推定する
 - `summary()` または `broom::tidy()` で結果を読む
 - `ggplot2::ggplot()` または `coefplot::coefplot()` で推定結果を可視化する

回帰分析

因果効果の推定のために

記号の設定

- 個体 $i = 1, 2, \dots, N$
- 結果変数（応答変数） Y_i
- 処置変数（説明変数） D_i
- 処置変数以外の説明変数（コントロール変数）
 $X_{1i}, X_{2i}, \dots, X_{ki}$

期待値 (expectation)*

- Y_i が連続型確率変数で確率密度関数が $f(y)$ で表されるとき、 Y の期待値 $\mathbb{E}[Y_i]$ は

$$\mathbb{E}[Y_i] = \int y f(y) dy$$

- Y が離散型確率変数のとき、 Y の期待値 $\mathbb{E}[Y_i]$ は

$$\mathbb{E}[Y_i] = \sum_y y \Pr(Y_i = y)$$

条件付き期待値*

- $X_i = x$ に条件付けた Y の期待値 $\mathbb{E}[Y_i | X_i = x]$ は

- ▶ Y が連続型変数のとき：

$$\mathbb{E}[Y_i | X_i = x] = \int y f(y | X_i = x) dy$$

- ▶ Y が離散型変数のとき：

$$\mathbb{E}[Y_i | X_i = x] = \sum_y y \Pr(Y_i = y | X_i = x)$$

- $\mathbb{E}[Y_i | X_i]$ は X の関数

繰り返し期待値の法則*

- $\mathbb{E} [\mathbb{E}[Y_i | X_i]] = \mathbb{E}[Y_i]$

▶ 離散の場合の証明 (連続の場合も同様に証明できる)

$$\begin{aligned}\mathbb{E} [\mathbb{E}[Y_i | X_i]] &= \mathbb{E} \left[\sum_y y \Pr(Y_i = y | X_i) \right] \\&= \sum_x \left[\sum_y y \Pr(Y_i = y | X_i = x) \right] \Pr(X_i = x) \\&= \sum_x \sum_y y \Pr(Y_i = y | X_i = x) \Pr(X_i = x) \\&= \sum_y y \left[\sum_x \Pr(Y_i = y, X_i = x) \right] \\&= \sum_y y \Pr(Y_i = y) = \mathbb{E}[Y_i].\end{aligned}$$

回帰 (regression)

- 結果変数の確率 [密度] を説明変数の関数で表す

$$p(Y | D, X_1, X_2, \dots, X_k) = f(D, X_1, X_2, \dots, X_k)$$

- 結果変数 Y を説明変数に回帰する

▶ 回帰関数： $\mathbb{E}[Y | D, X_1, X_2, \dots, X_k]$

- 回帰関数は、説明変数（処置およびコントロール）で条件付けた Y の条件付き期待値

▶ 回帰関数が線形関数だと**仮定**すると

$$\mathbb{E}[Y_i | D_i, X_1, X_2, \dots, X_k] = \alpha + \beta D_i + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \dots \gamma_k X_{ki}$$

単回帰 (simple regression)

- Y を D に回帰する
 - ▶ 回帰関数： $\mathbb{E}[Y \mid D]$
 - 回帰関数は、説明変数 D で条件付けた Y の条件付き期待値
 - ▶ 回帰関数が線形関数だと**仮定**すると

$$\mathbb{E}[Y_i \mid D_i] = \alpha + \beta D_i$$

観測値は回帰関数と残差で構成される

$$Y_i | D_i = \mathbb{E}[Y_i | D_i] + (Y_i | D_i - \mathbb{E}[Y_i | D_i])$$

▶ 残差 : $e_i | D_i = Y_i | D_i - \mathbb{E}[Y_i | D_i]$

– $\mathbb{E}[e_i] = 0$

– $\text{Cov}(D_i, e_i) = \mathbb{E}[D_i e_i] = 0$

$$e_i = Y_i | D_i - \mathbb{E}[Y_i | D_i]$$

$$Y_i | D_i = \mathbb{E}[Y_i | D_i] + e_i = \alpha + \beta D_i + e_i$$

「傾き」は条件付き期待値の差 (1)

$$Y_i | D_i = \alpha + \beta D_i + e_i$$

- 処置が二値変数のとき : $D_i \in \{0,1\}$
 - ▶ $\mathbb{E}[Y_i | D_i = 0] = \mathbb{E}[\alpha + \beta \cdot 0 + e_i] = \alpha$
 - ▶ $\mathbb{E}[Y_i | D_i = 1] = \mathbb{E}[\alpha + \beta \cdot 1 + e_i] = \alpha + \beta$
- $\mathbb{E}[Y_i | D_i = 1] - \mathbb{E}[Y_i | D_i = 0] = \beta$
 - ▶ β : 処置 D の値が0から1に変わったとき、結果変数 Y の期待値がどれだけ増えるかを表す

「傾き」は条件付き期待値の差 (2)

$$Y_i | D_i = \alpha + \beta D_i + e_i$$

- 処置 D_i が二値変数ではないとき : $D_i \in \mathbb{R}$
 - ▶ $\mathbb{E}[Y_i | D_i = d] = \mathbb{E}[\alpha + \beta \cdot d + e_i] = \alpha + \beta d$
 - ▶ $\mathbb{E}[Y_i | D_i = d + 1] = \mathbb{E}[\alpha + \beta \cdot (d + 1) + e_i] = \alpha + \beta d + \beta$
- $\mathbb{E}[Y_i | D_i = d + 1] - \mathbb{E}[Y_i | D_i = d] = \beta$
 - ▶ β : 処置変数 D の値が1単位分増えたとき、結果変数 Y の期待値がどれだけ増えるかを表す

因果効果と回帰直線の「傾き」(1)

$$Y_i | D_i = \alpha + \beta D_i + e_i$$

- 処置が二値変数のとき : $D_i \in \{0,1\}$

$$\begin{aligned}\beta &= \mathbb{E}[Y_i | D_i = 1] - \mathbb{E}[Y_i | D_i = 0] \\ &= \mathbb{E}[Y_i(1) | D_i = 1] - \mathbb{E}[Y_i(0) | D_i = 0]\end{aligned}$$

- ▶ 回帰直線の傾き : 処置群と統制群の観測された平均値の差

因果効果と回帰直線の「傾き」(2)

- 観測された平均値の差：

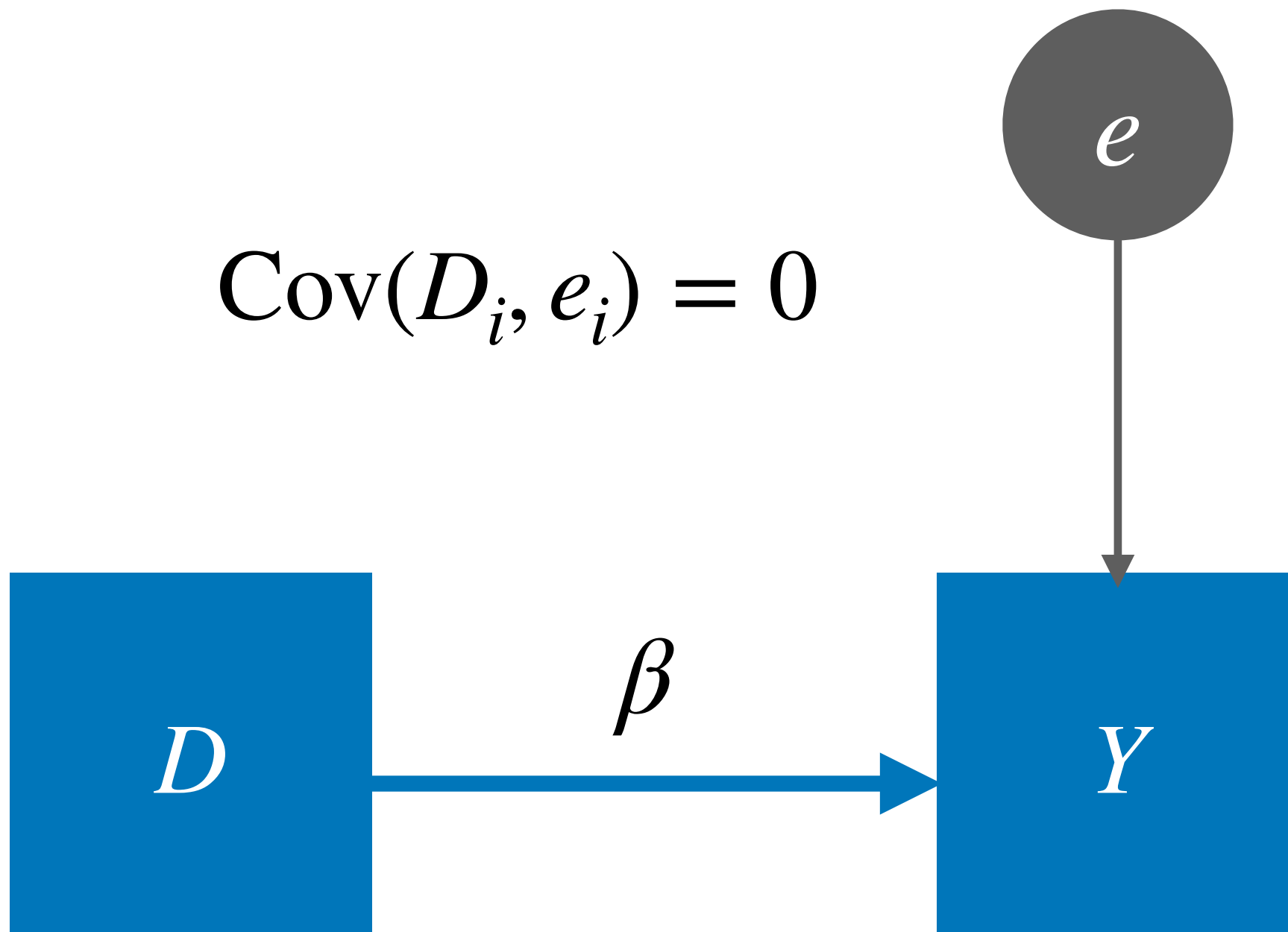
$$\begin{aligned}\beta &= \mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0] \\ &= \mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 1] \\ &\quad + \mathbb{E}[Y_i(0) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0] \\ &= \text{ATT} + \text{セレクションバイアス}\end{aligned}$$

- ▶ セレクションバイアスが 0 なら： $\beta = \text{ATT}$
- ▶ 平均独立が成り立つなら：

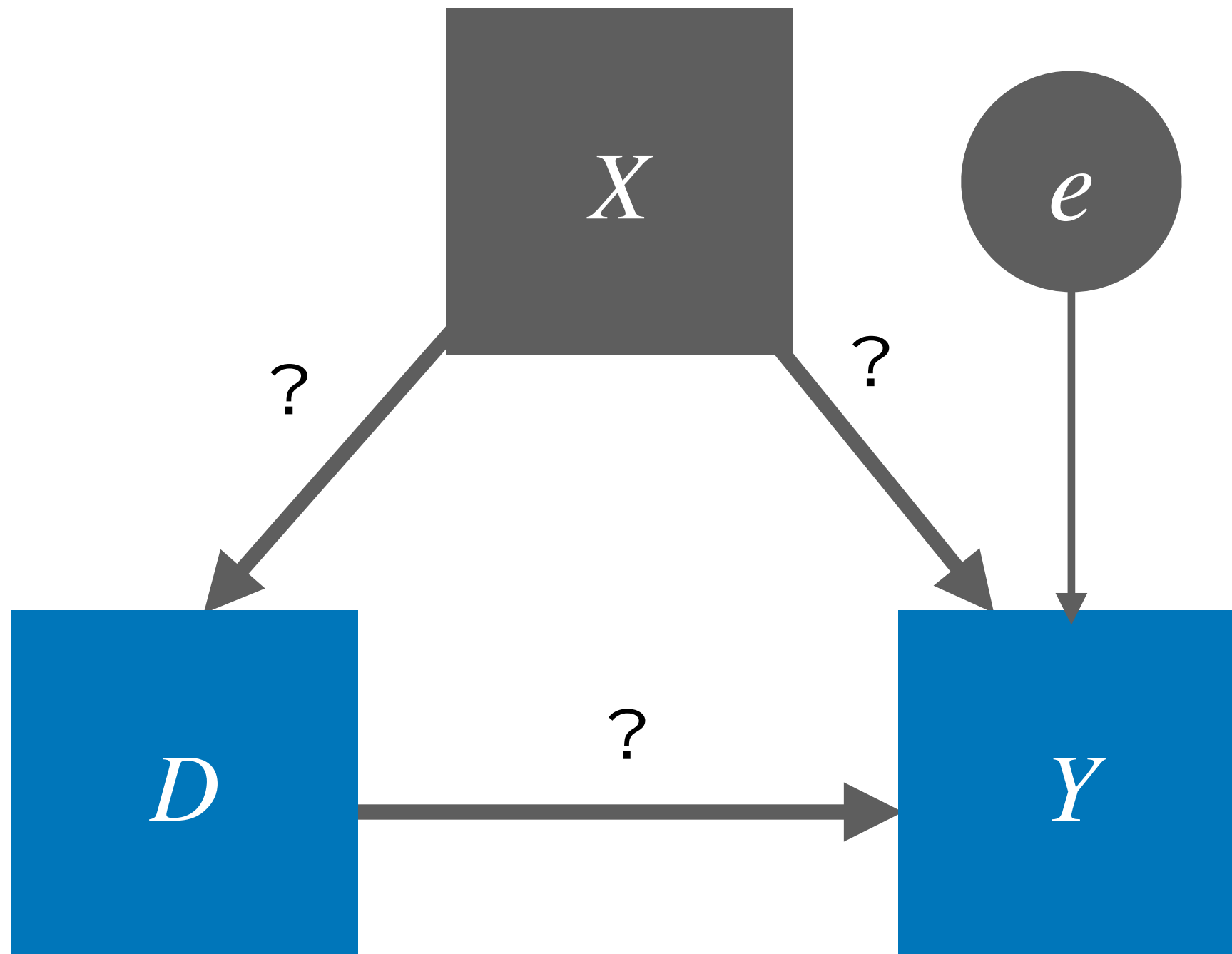
$$\begin{aligned}\beta &= \mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0] \\ &= \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \\ &= \text{ATE}\end{aligned}$$

ここで考えている関係

$$\text{Cov}(D_i, e_i) = 0$$



セルフセレクションがあったら？



重回帰 (multiple regression)

- セレクションを考慮に入れた回帰式を作る
 - ▶ Y は D と X の関数
 - 回帰関数： D と X で条件付けた Y の期待値

$$\mathbb{E}[Y_i | D_i, X_i] = \alpha + \beta D_i + \gamma X_i$$

$$Y_i | D_i, X_i = \mathbb{E}[Y_i | D_i, X_i] + e_i = \alpha + \beta D_i + \gamma X_i + e_i$$

「傾き」は条件付き期待値の差 (3)

$$Y_i | D_i, X_i = \alpha + \beta D_i + \gamma X_i + e_i$$

- 処置が二値変数のとき : $D_i \in \{0,1\}$
 - ▶ $\mathbb{E}[Y_i | D_i = 0, X_i = x] = \mathbb{E}[\alpha + \beta \cdot 0 + \gamma x + e_i] = \alpha + \gamma x$
 - ▶ $\mathbb{E}[Y_i | D_i = 1, X_i = x] = \mathbb{E}[\alpha + \beta \cdot 1 + \gamma x + e_i] = \alpha + \beta + \gamma x$
- $\mathbb{E}[Y_i | D_i = 1, X_i = x] - \mathbb{E}[Y_i | D_i = 0, X_i = x] = \beta$
 - ▶ β : $X = x$ のとき、処置 D の値が0から1に変わると結果変数 Y の期待値はどれだけ増えるかを表す

因果効果と重回帰における「傾き」

$$\begin{aligned}\beta &= \mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x] \\ &= \mathbb{E}[Y_i(1) \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i(0) \mid D_i = 0, X_i = x]\end{aligned}$$

ここで

$$\begin{cases} \mathbb{E}[Y_i(1) \mid D_i = 1, X_i = x] = \mathbb{E}[Y_i(1) \mid D_i = 0, X_i = x] \\ \text{and} \\ \mathbb{E}[Y_i(0) \mid D_i = 1, X_i = x] = \mathbb{E}[Y_i(0) \mid D_i = 0, X_i = x] \end{cases}$$

が成り立つなら、

$$\begin{aligned}\beta &= \mathbb{E}[Y_i(1) \mid X_i = x] - \mathbb{E}[Y_i(0) \mid X_i = x] \\ &= \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]\end{aligned}$$

▶ 回帰係数 β : X で条件付けた ATE

条件付き平均独立

- **条件付き平均独立** (conditional mean independence; conditional mean exchangeability)

$$\mathbb{E}[Y_i(1) \mid D_i = 1, X] = \mathbb{E}[Y_i(1) \mid D_i = 0, X]$$

かつ

$$\mathbb{E}[Y_i(0) \mid D_i = 1, X] = \mathbb{E}[Y_i(0) \mid D_i = 0, X]$$

- 条件付き平均独立が成り立つとき：

$$\mathbb{E}[Y_i \mid D_i = 1, X_i] - \mathbb{E}[Y_i \mid D_i = 0, X_i]$$

$$= \mathbb{E}[Y_i(1) \mid D_i = 1, X_i] - \mathbb{E}[Y_i(0) \mid D_i = 0, X_i]$$

$$= \mathbb{E}[Y_i(1) \mid X_i] - \mathbb{E}[Y_i(0) \mid X_i]$$

$$\mathbb{E} \left(\mathbb{E}[Y_i(1) \mid X_i] - \mathbb{E}[Y_i(0) \mid X_i] \right) = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] = \text{ATE}$$

条件付き独立・条件付き交換可能性

- ・セレクションバイアスの原因が X だけで、 X の影響さえ取り除けば、 D の値はランダムに決まると**仮定**すると：潜在的結果と処置は、

- ▶ 条件付き独立： $\{Y(0), Y(1)\} \perp\!\!\!\perp D \mid X$

- ▶ 条件付き交換可能性：

$$p(Y(0), Y(1) \mid D = 1, X) = p(Y(0), Y(1) \mid D = 0, X) = p(Y(0), Y(1) \mid X)$$

- ・セレクションバイアスの原因が X_1, X_2, \dots, X_k なら、

- ▶ 条件付き独立： $\{Y(0), Y(1)\} \perp\!\!\!\perp D \mid X_1, X_2, \dots, X_k$

- ・条件付き独立 \Rightarrow 条件付き平均独立

- ・調査・観察研究の問題：セレクションバイアスの原因を全て特定し、観察するのが難しい

無視可能性 (ignorability)*

- 強い意味での無視可能性 (strong ignobility) の仮定：観測された共変量に条件付ければ、潜在変数と処置の割付けは独立

$$p(D \mid Y(0), Y(1), X) = p(D \mid X) \quad \text{強い意味での無視可能性}$$

$$\Leftrightarrow p(Y(0), Y(1) \mid D, X) = p(Y(0), Y(1) \mid X) \quad \text{条件付き交換可能性}$$

$$\Rightarrow p(Y(D) \mid D, X) = (Y(D) \mid X) \quad (D = 0, 1) \quad \text{弱い無視可能性}$$

- ▶ 処置の割付けは観測された変数だけに依存する
(selection on observables) という仮定

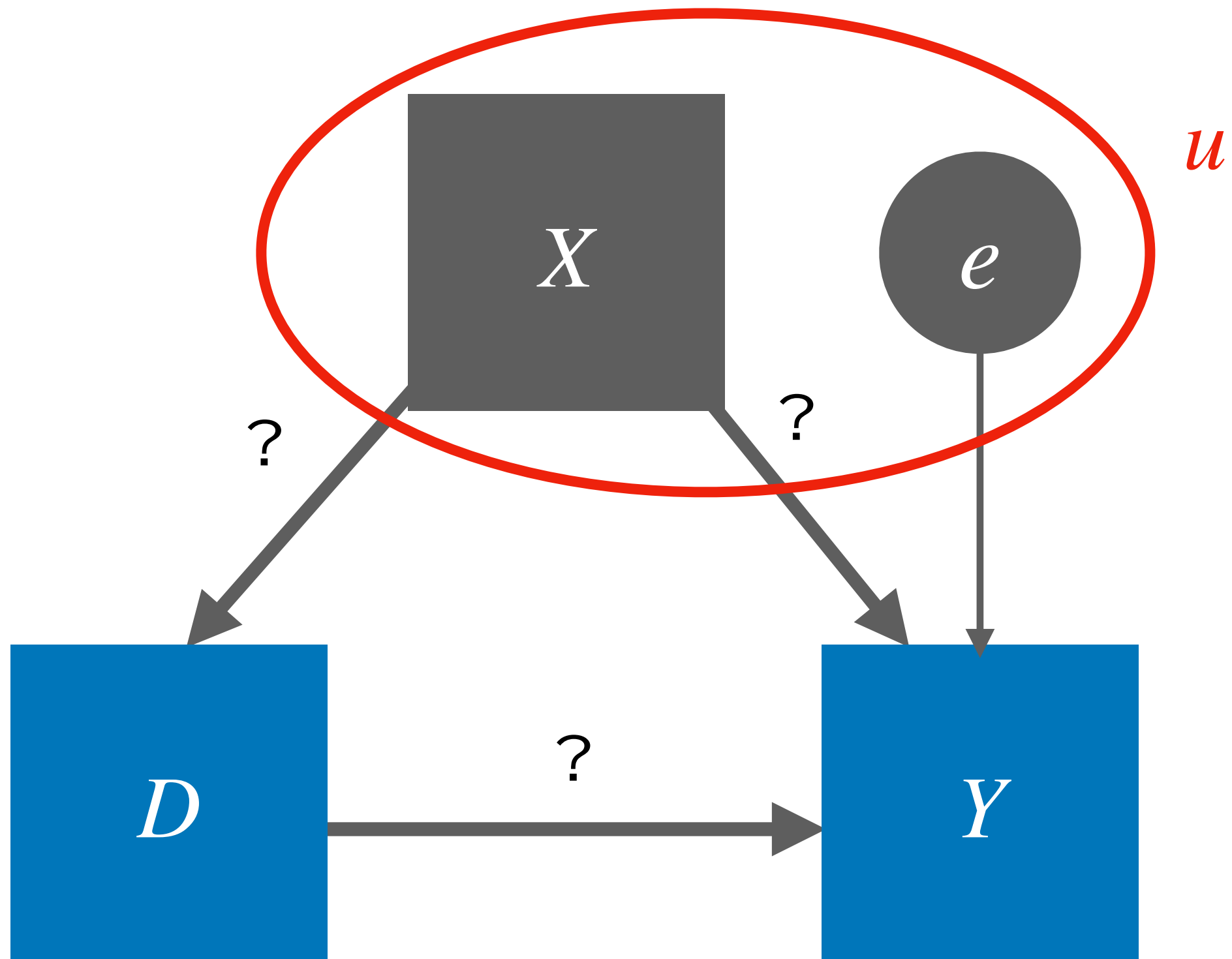
セレクションと重回帰

- セレクションバイアスがありそうな調査・観察データでも、重回帰によってATEを推定できる
- そのためには、以下の2つが必要
 - ▶ セレクションを生み出す変数を**観測**する
 - ▶ セレクションを生み出す変数を回帰式に含める
- これができれば、セレクションバイアスは除去できる
 - ▶ 完全にできない場合、セレクションバイアスをゼロにすることはできないが、減らすことはできる
- セレクションバイアスを生み出す変数：**交絡因子（共変量）**

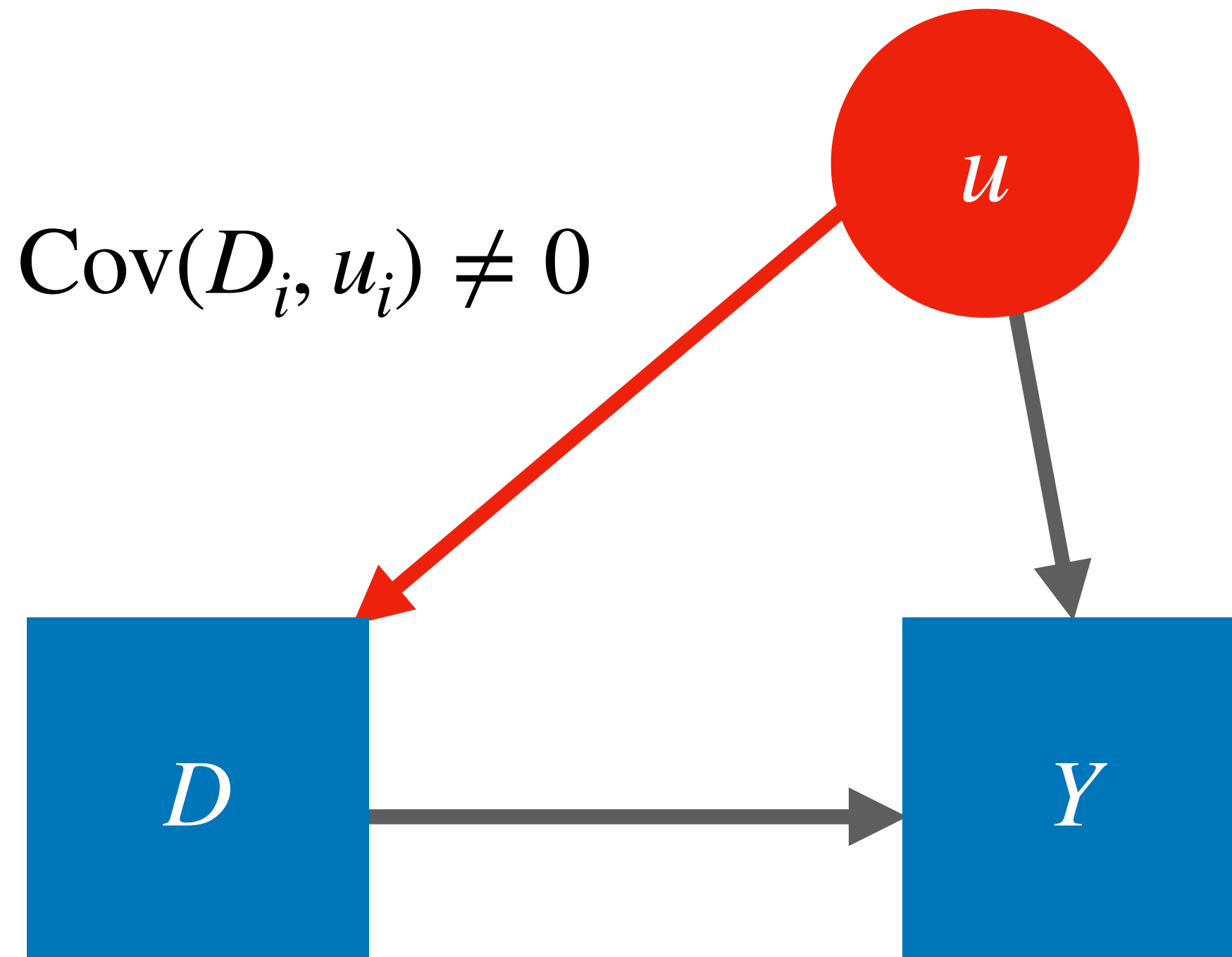
回帰分析のバイアス I

脱落変数バイアス

セレクションバイアスがあったら？



セレクションバイアスがあったら？



回帰モデルの定式化

- セレクションは X によって生じると仮定する

- ▶ 正しい定式化 (long regression)

$$Y_i = \alpha^l + \beta^l D_i + \gamma^l X_i + e_i \quad (1)$$

- ▶ セレクションを考慮しない定式化 (short regression)

$$Y_i = \alpha^s + \beta^s D_i + u_i \quad (2)$$

- ▶ X を D に回帰する

$$X_i = \alpha_0 + \lambda D_i + \nu_i \quad (3)$$

セレクションを無視する

- 正しい式から X を消去する
- (1) に (3) を代入する

$$Y_i = \alpha^l + \beta^l D_i + \gamma^l X_i + e_i$$

$$= \alpha^l + \beta^l D_i + \gamma^l (\alpha_0 + \lambda D_i + \nu_i) + e_i$$

$$= \alpha^l + \gamma^l \alpha_0 + (\beta^l + \gamma^l \lambda) D_i + e_i + \gamma^l \nu_i \quad (4)$$

脱落変数バイアス (OVB)

- 脱落 [欠落] 変数バイアス : omitted variable bias
- 式(2) と (4) : 式 (1) から X_i が脱落している
 - ▶ Y を D に回帰したときの D の係数 :
 - $\beta^s = \beta^l + \gamma^l \lambda$
 - 脱落変数バイアス : $\gamma^l \lambda$
 - ◆ γ^l : X と Y の共変関係
 - ◆ λ : X と D の共変関係

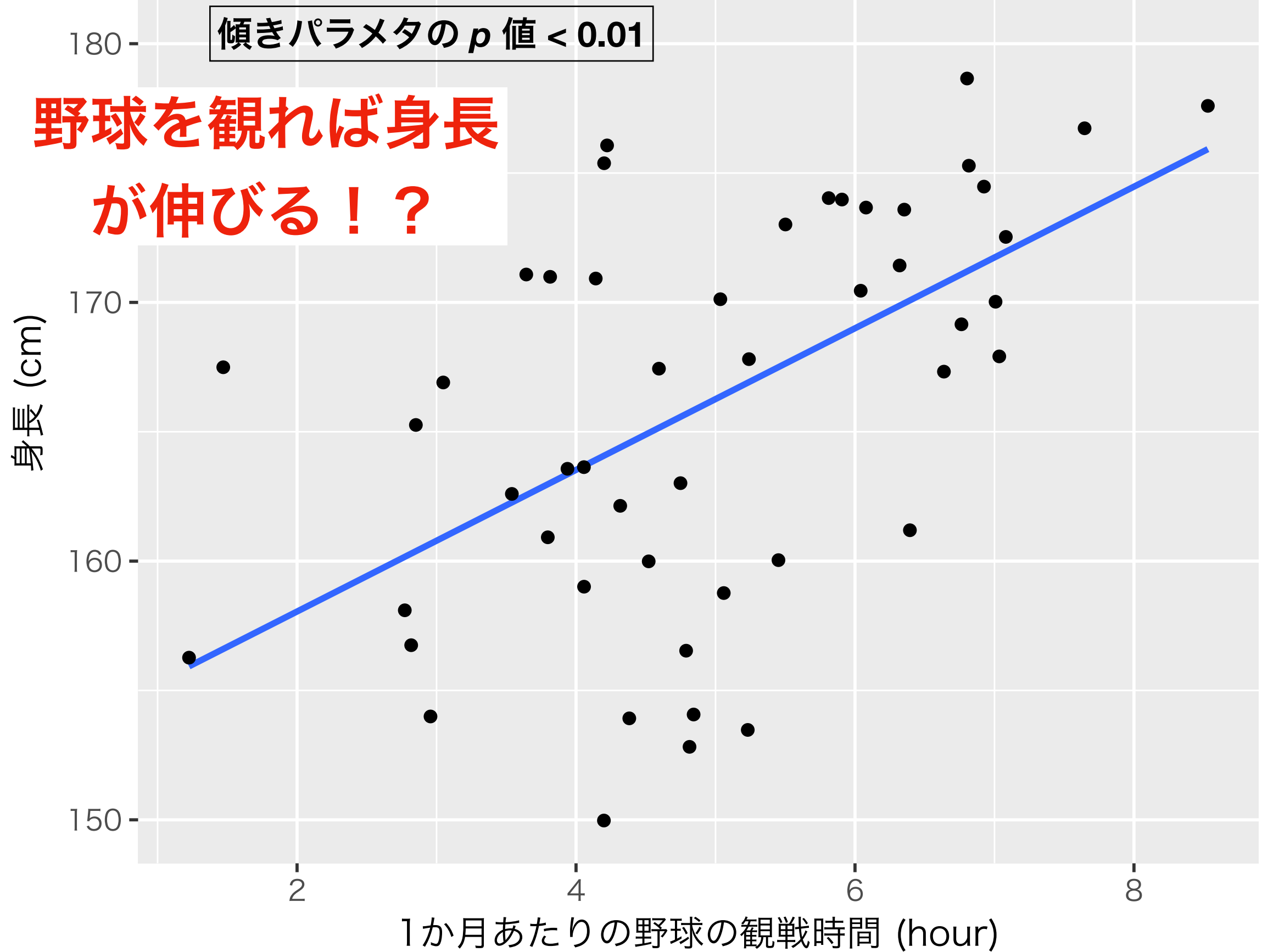
脱落変数バイアスと交絡

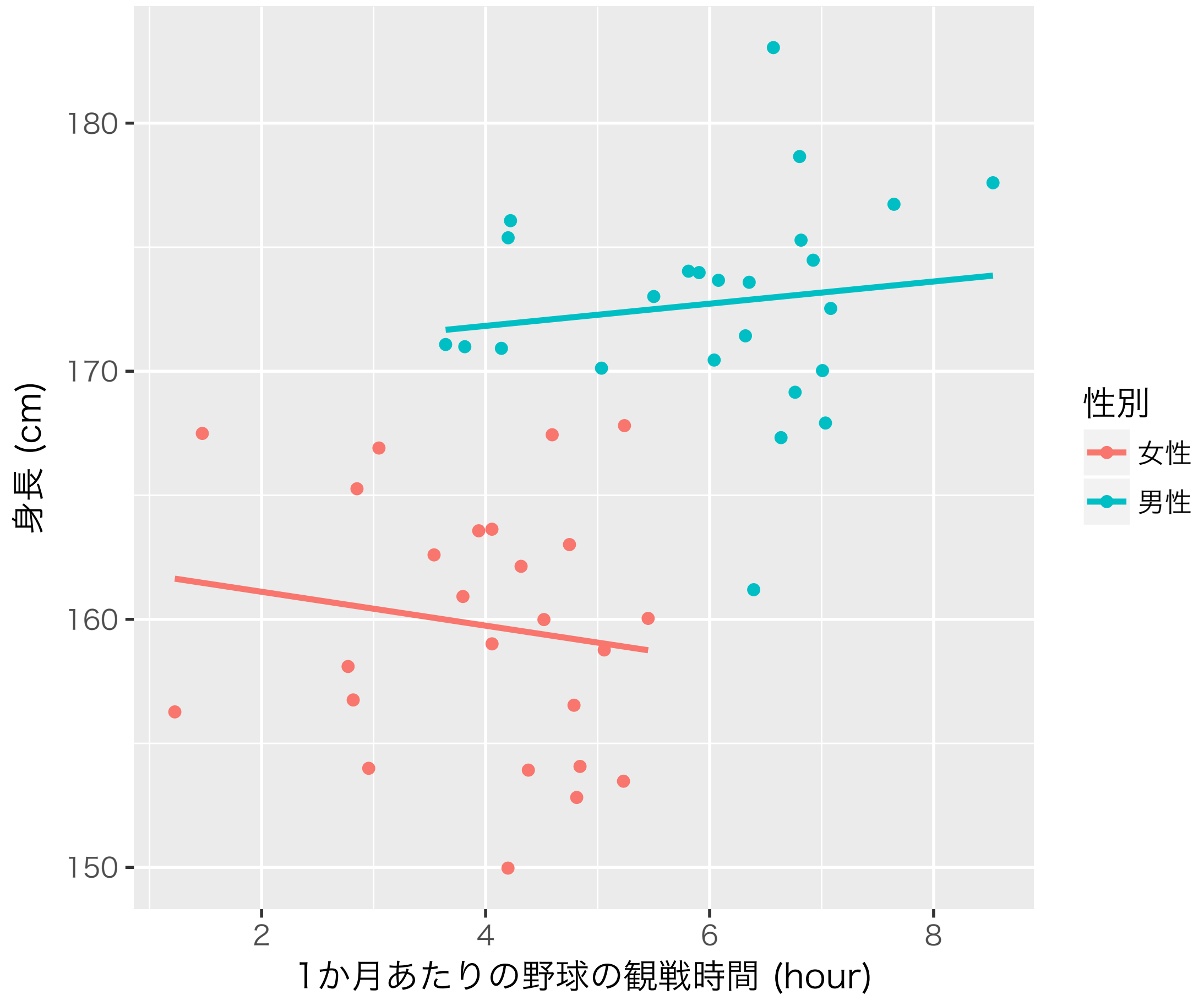
- 脱落変数バイアス： $\gamma^l \lambda$
 - ▶ $\gamma^l = 0$ または $\lambda = 0$ ならば、このバイアスは生じない
 - ▶ $\gamma^l \neq 0$ かつ $\lambda \neq 0$ のとき、 X を 交絡因子（共変量）と呼ぶ
- 交絡をコントロールしないと
 - ▶ 脱落変数バイアスが生じる
 - ▶ つまり、セレクションバイアスが除去されずに残る

脱落変数バイアスの例

- 身長とプロ野球の観戦時間の関係は？
 - ▶ プロ野球の観戦時間は身長を伸ばす？
 - ▶ 理論的に考えると、おそらく No!
 - ▶ しかし、回帰分析をすると…
 - Yes ???

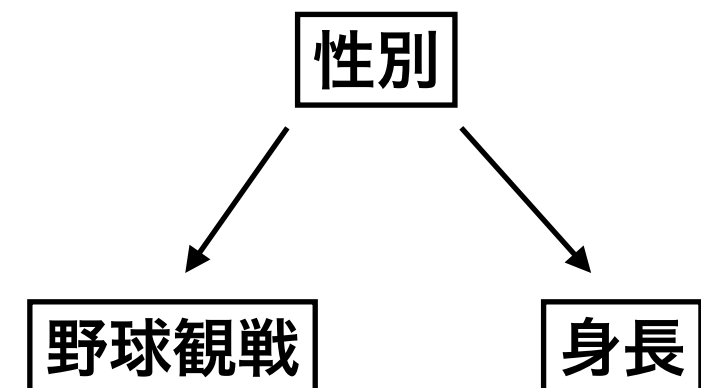
(架空のデータ)





何が問題か？

- 性別が交絡になっている
- 性別が野球の観戦時間 (X) と身長 (Y) の両者に影響を及ぼす
 - ▶ 男性のほうが野球を観る
 - ▶ 男性のほうが身長が高い



男性型脱毛症と新型コロナウイルス

- 男性型脱毛症 [Androgenetic Alopecia]（あるいはその原因となるホルモン [androgen]）は、新型コロナウイルスの重症化リスクを高める! (???)
 - ▶ Wambier et al. 2020. “Androgenetic Alopecia Presents in the Majority of Hospitalized COVID-19 Patients,” <https://doi.org/10.1016/j.jaad.2020.05.079>
 - ▶ Goren et al. 2020. “A Preliminary Observation: Male Pattern in Hair Loss among Hospitalized COVID-19 Patients in Spain”, <https://doi.org/10.1111/jocd.13443>
- 因果効果は疑わしい
 - ▶ 年齢がコントロールされていない！
 - 参考：<https://www.forbes.com/sites/marlamilling/2020/06/06/bald-men-at-higher-risk-of-severe-coronavirus-symptoms/#2449f87729e4>

回帰分析におけるコントロール

- コントロール変数
 - ▶ RCT におけるブロック変数の役割を果たす
- 重回帰がやっていること
 - ▶ コントロール変数によるブロッキング
 - ▶ ブロックごとに処置効果を計算
 - ▶ ブロックごとの処置効果の加重平均を計算
 - $X = x$ となるブロックの重み
 - ◆ ATE: $\Pr(X_i = x)$,
 - ◆ ATT: $\Pr(X_i = x \mid D_i = 1)$
- ❖ 詳しくは、Angrist and Pischke (2008) 3.3.1 節を参照

コントロール変数による条件付け

- 交絡因子 X を統制（コントロールする）
- 交絡因子は複数あることも: X_1, X_2, \dots, X_k
 - ▶ 私たちが比較したい個体が様々な面で異質なとき、複数の交絡を統制する必要がある
- 複数の交絡を統制するためには、標本サイズはある程度大きくないといけない
 - ▶ 標本サイズが小さいと、各ブロックに属する個体数が少なくなる
 - ▶ 処置の値が異なる個体が存在しないブロックの重みはゼロ

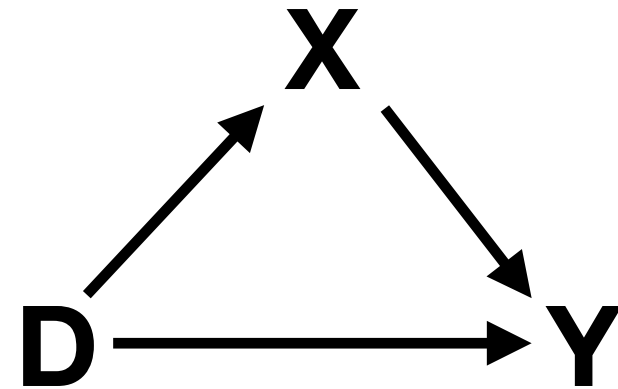
回帰分析のバイアス II

処置後変数バイアス

処置の影響を受けた変数を含む重回帰

- Y を D と X に回帰する

$$Y_i = \alpha + \beta D_i + \gamma X_i + e_i$$



- ▶ D が Y に与える処置効果を知りたいわけではないなら、何も問題ない
 - X が Y に与える影響を知りたいなら、正しい推定
- しかし、 D が Y に与える処置効果を知りたいなら、この回帰式は問題

処置後変数

- X は D の処置後変数 (post-treatment variable)
 - ▶ D の処置効果の一部は、 X を通じて Y に伝わる

- $X_i = \alpha_0 + \lambda D_i + u_i$

- これを先程の式に代入し、 X を消去する

$$Y_i = \alpha + \beta D_i + \gamma X_i + e_i$$

$$= \alpha + \beta D_i + \gamma(\alpha_0 + \lambda D_i + u_i) + e_i$$

$$= (\alpha + \alpha_0) + (\beta + \gamma\lambda)D_i + (\gamma u_i + e_i)$$

処置後変数バイアス

- Y を D と X に回帰したときの推定値： β
- Y を D のみに回帰したときの推定値： $\beta + \gamma\lambda$
 - ▶ これが、 D の Y に対する処置効果
- 処置後変数によって生じたバイアス： $-\gamma\lambda$
 - ▶ γ と λ の符号が同じ：バイアスにより過小推定
 - ▶ γ と λ の符号が異なる：バイアスにより過大推定
 - ▶ γ または λ が0：バイアスは生じない
 - $\lambda = 0$ なら X は D の処置後変数ではない

重回帰における

コントロール変数の選び方

どの変数を統制する？

- 重回帰で因果推論を行うために使う変数は何？
 - ▶ 結果変数（理論における結果）：絶対に必要
 - ▶ 処置変数（理論における原因）：絶対に必要
 - ▶ 統制変数：必要かもしれない（ほとんどの場合必要）
 - どの変数を統制する？
 - いくつの変数を統制する？

バックドア基準

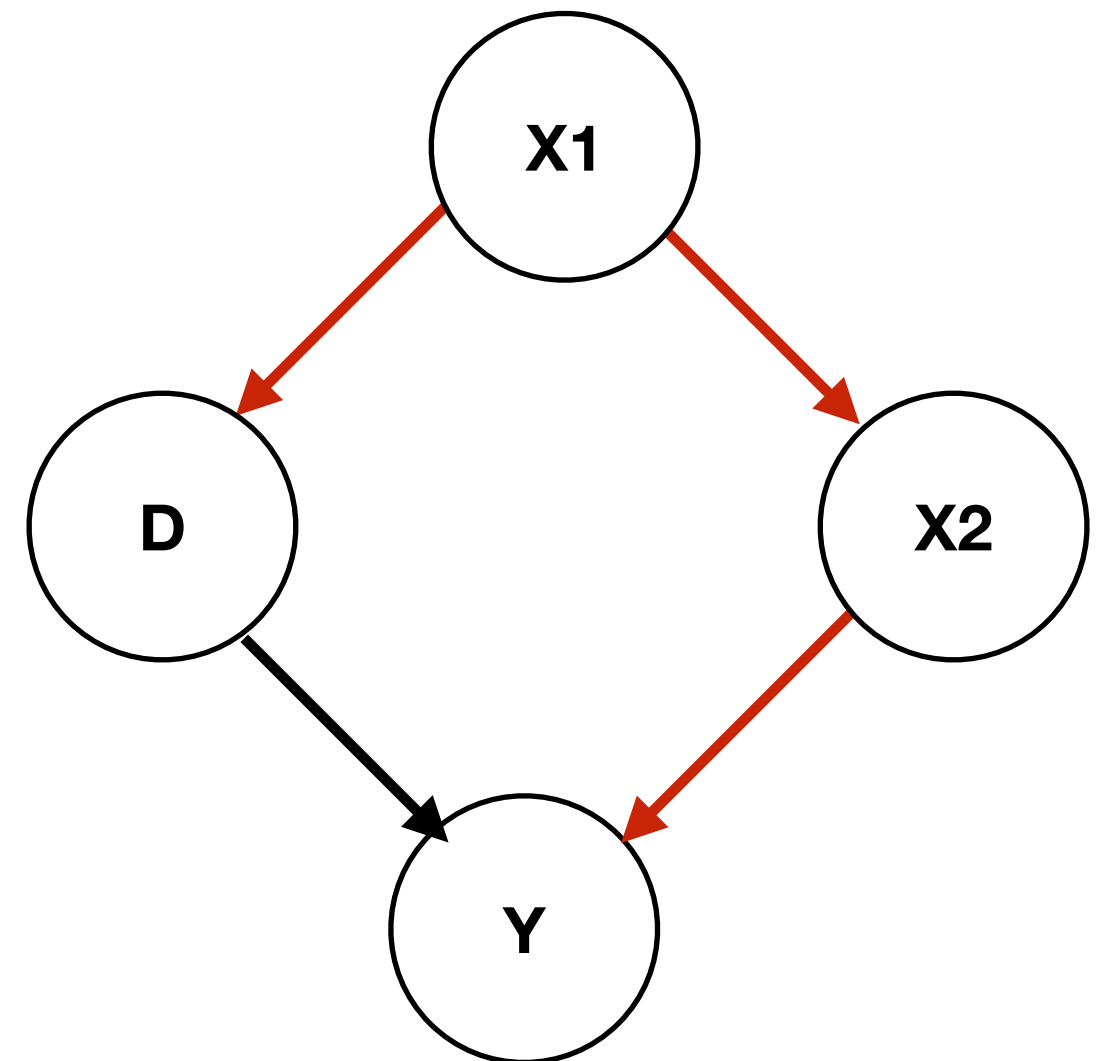
- どの変数を統制すべきか教えてくれる基準
- この用語は、因果推論におけるグラフィカルモデリングで使われる
 - ▶ DAG: directed acyclic graph、有向非巡回グラフ
 - ▶ 回帰分析でもこの考え方は便利
 - 詳しくは、以下を参照
 - ◆ 黒木学, 2017, 『構造的因果モデルの基礎』 共立出版.
 - ◆ Pearl, J. et al. (落海 訳) 2019, 『入門 統計的因果推論』 朝倉書店.

バックドア基準の基礎

- D : 処置変数 [treatment] (介入、刺激、暴露 [exposure]、独立変数)
- Y : 結果変数 [outcome] (応答変数、目的変数、従属変数)
- X : 統制変数 (コントロール、**交絡 [confounder]**、**共変量 [covariate]**)

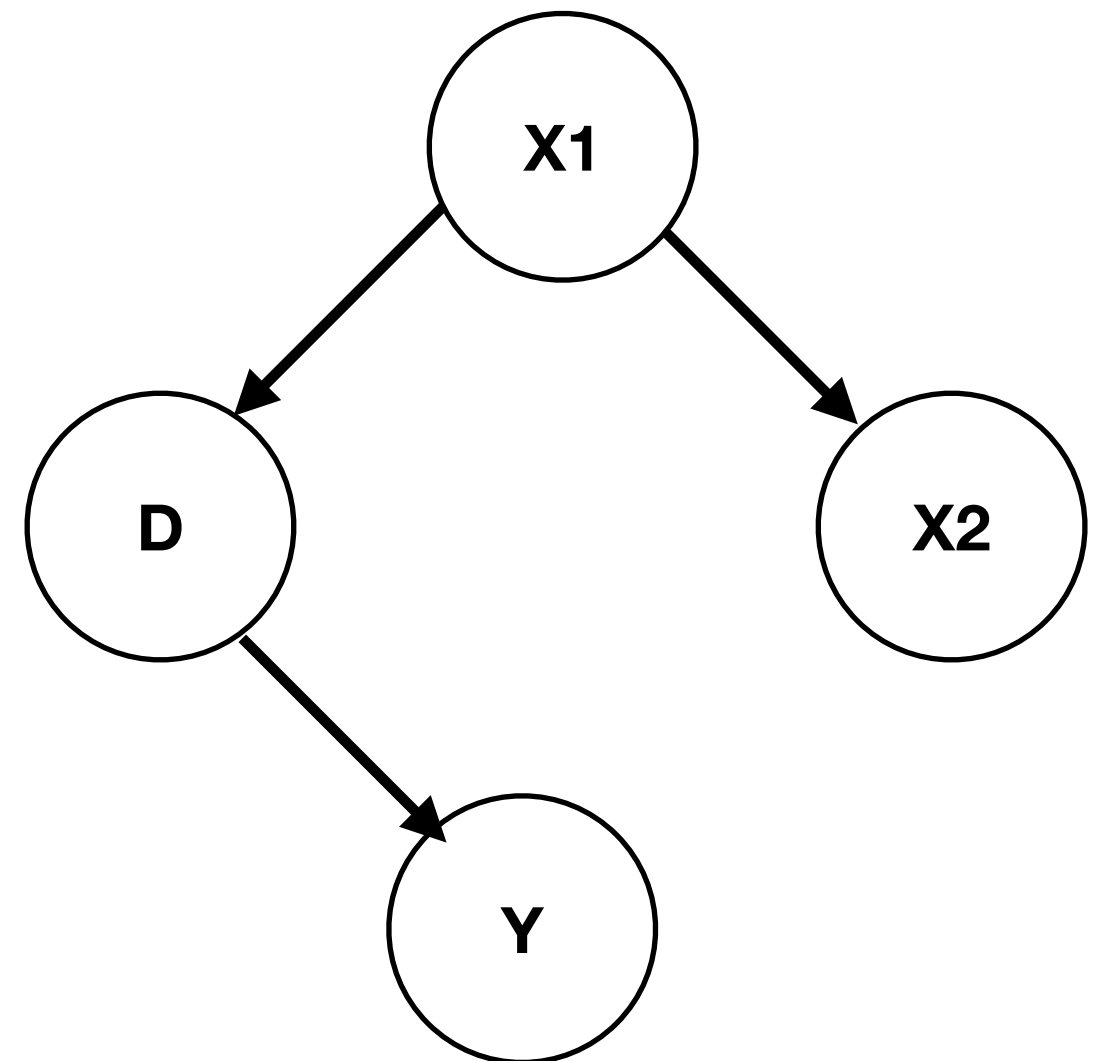
交絡変数とバックドア経路 (1)

- DAG を描いて考える
- バックドア経路: ある変数が**D**と**Y**の**両者**の原因となるような経路
 - ▶ $D \leftarrow X1 \rightarrow X2 \rightarrow Y$
- **交絡変数** (confounding variables, confounders): **D**と**Y**の**両者**の原因となる変数
 - ▶ $X1$



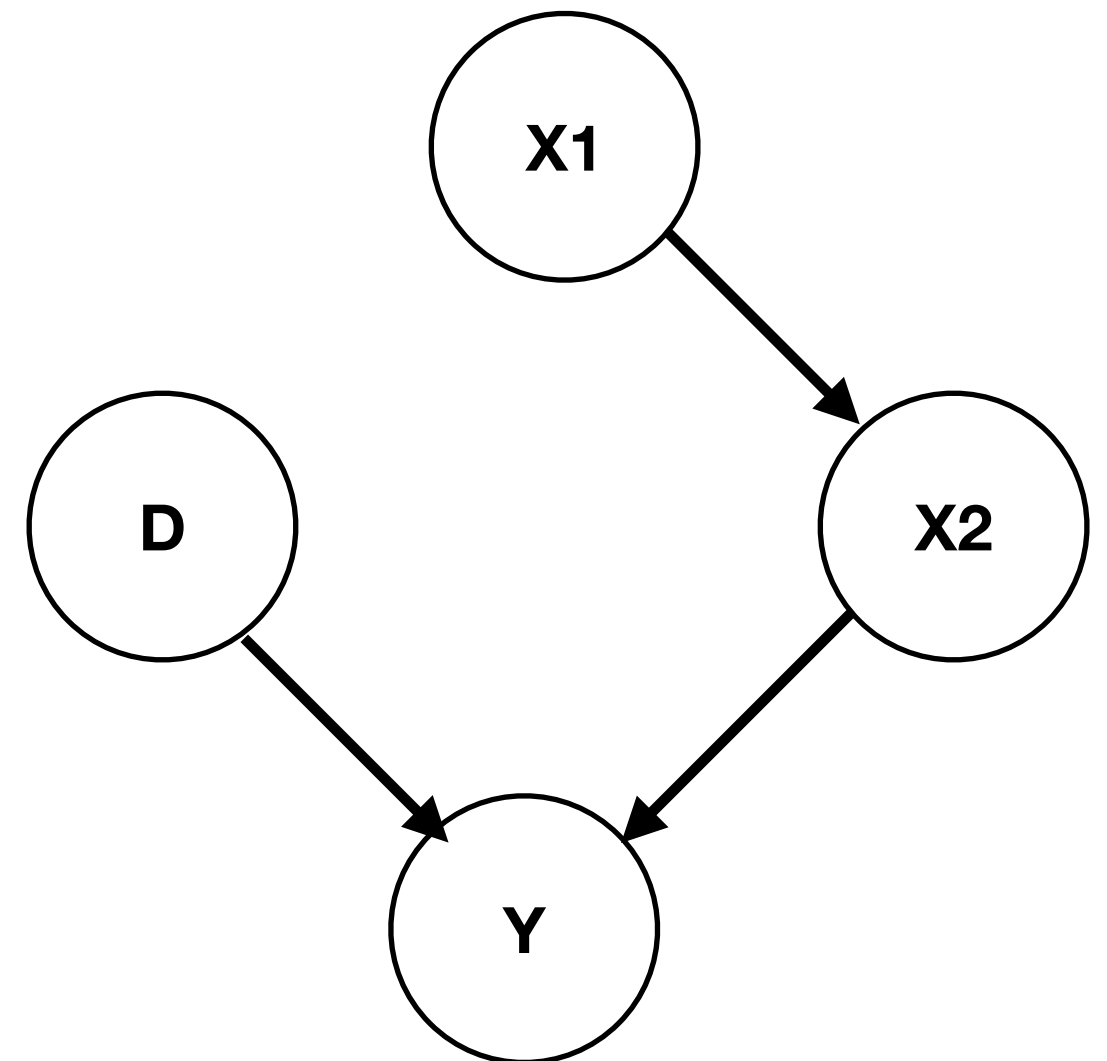
交絡変数とバックドア経路 (2)

- 右の図にバックドア経路は存在しない
 - ▶ $D \leftarrow X1 \rightarrow X2$ はバックドア経路ではない！
- 交絡変数はない
 - ▶ $X1$ は交絡ではない



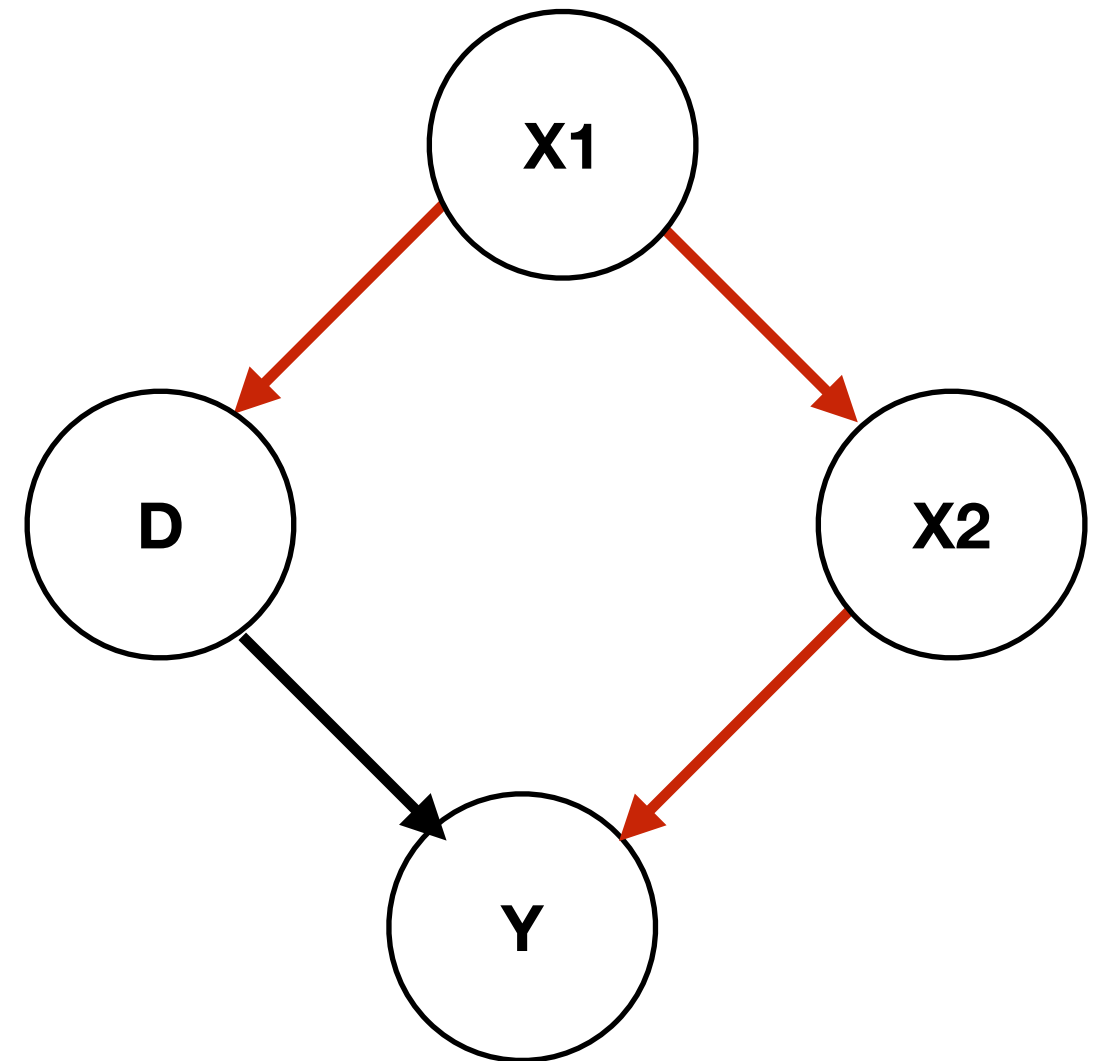
交絡変数とバックドア経路 (3)

- 右の図にバックドア経路は存在しない
 - ▶ $X1 \rightarrow X2 \rightarrow Y$ はバックドア経路ではない
- 交絡変数はない



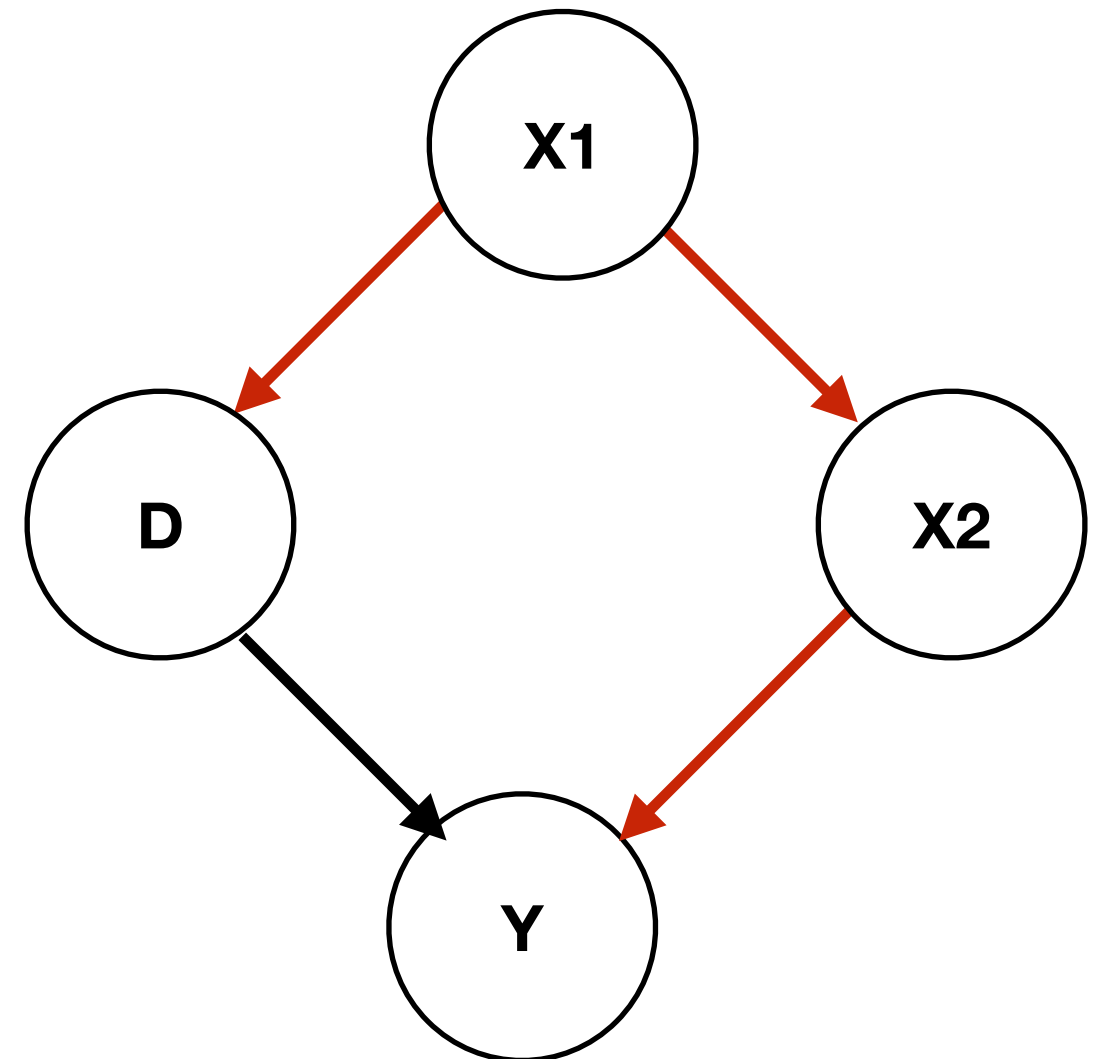
バックドアを閉じたい

- バックドア経路：
 - ▶ $D \leftarrow X1 \rightarrow X2 \rightarrow Y$
- バックドアを閉じたい
- どうすればいい？



バックドアを閉じる

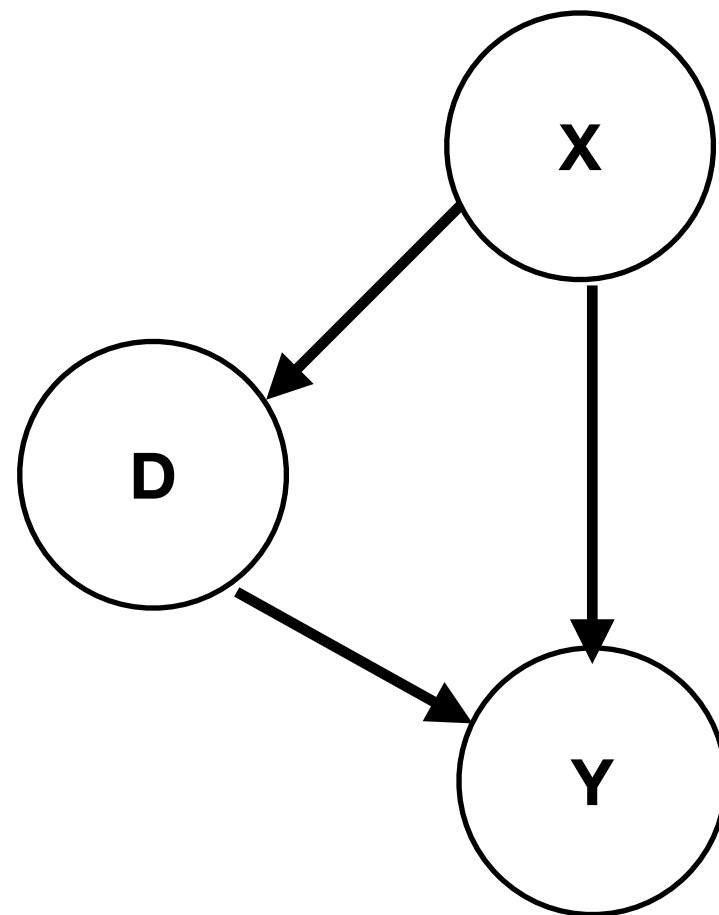
- バックドア経路にある変数をコントロールすれば良い！
- バックドア経路：
 - ▶ $D \leftarrow X1 \rightarrow X2 \rightarrow Y$
- この例では、閉じ方は3通り
 - ▶ $X2$ をコントロール
 - ▶ $X1$ をコントロール
 - ▶ $X1$ と $X2$ をコントロール



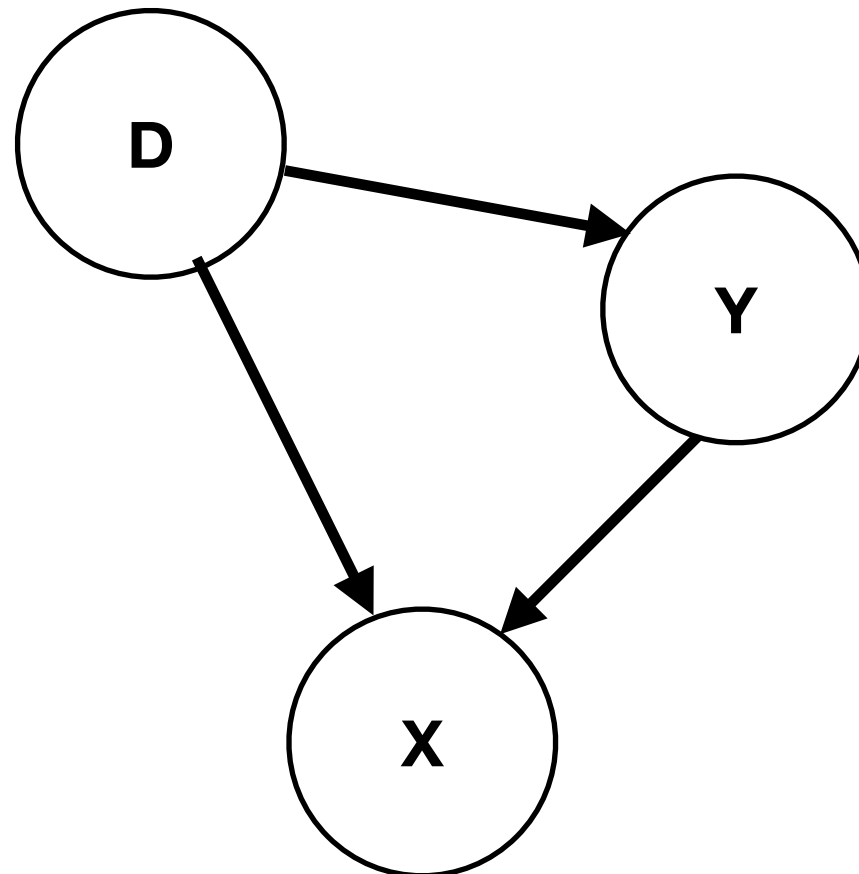
変数 D , Y , X の関係

- Y を結果、 D を原因とする
- 3つの可能性
 1. X は D と Y の交絡変数 (confounder) である
 2. X は D と Y の合流点 (collider) である
 3. X は D と Y の媒介変数 (mediator, 中間因子) である

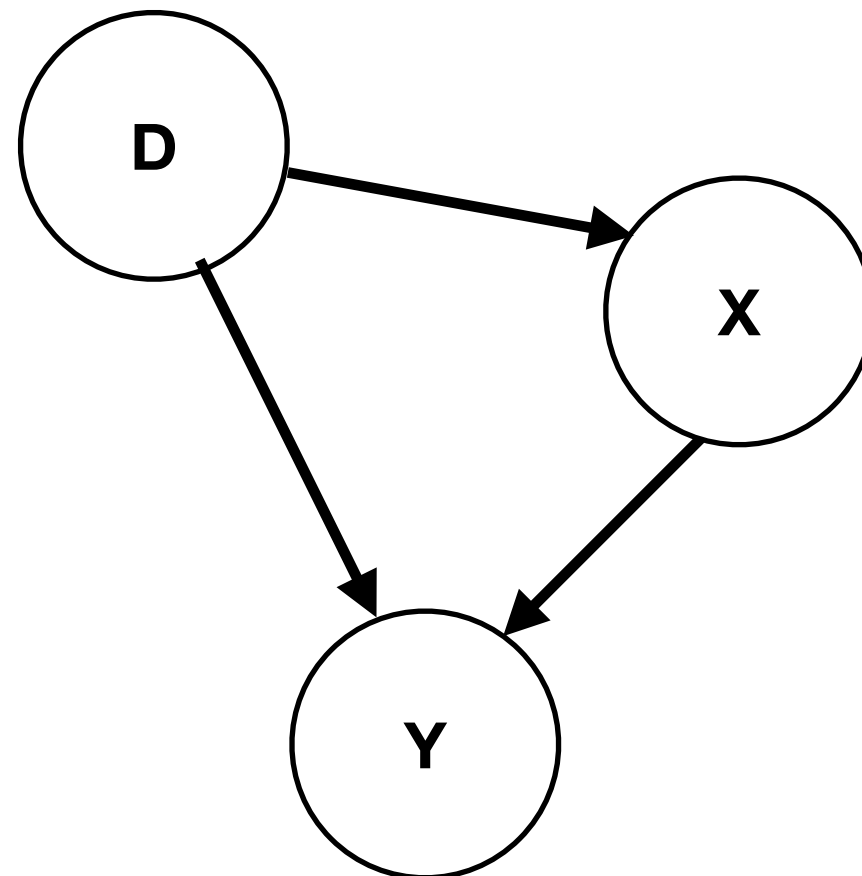
交絡変数 X



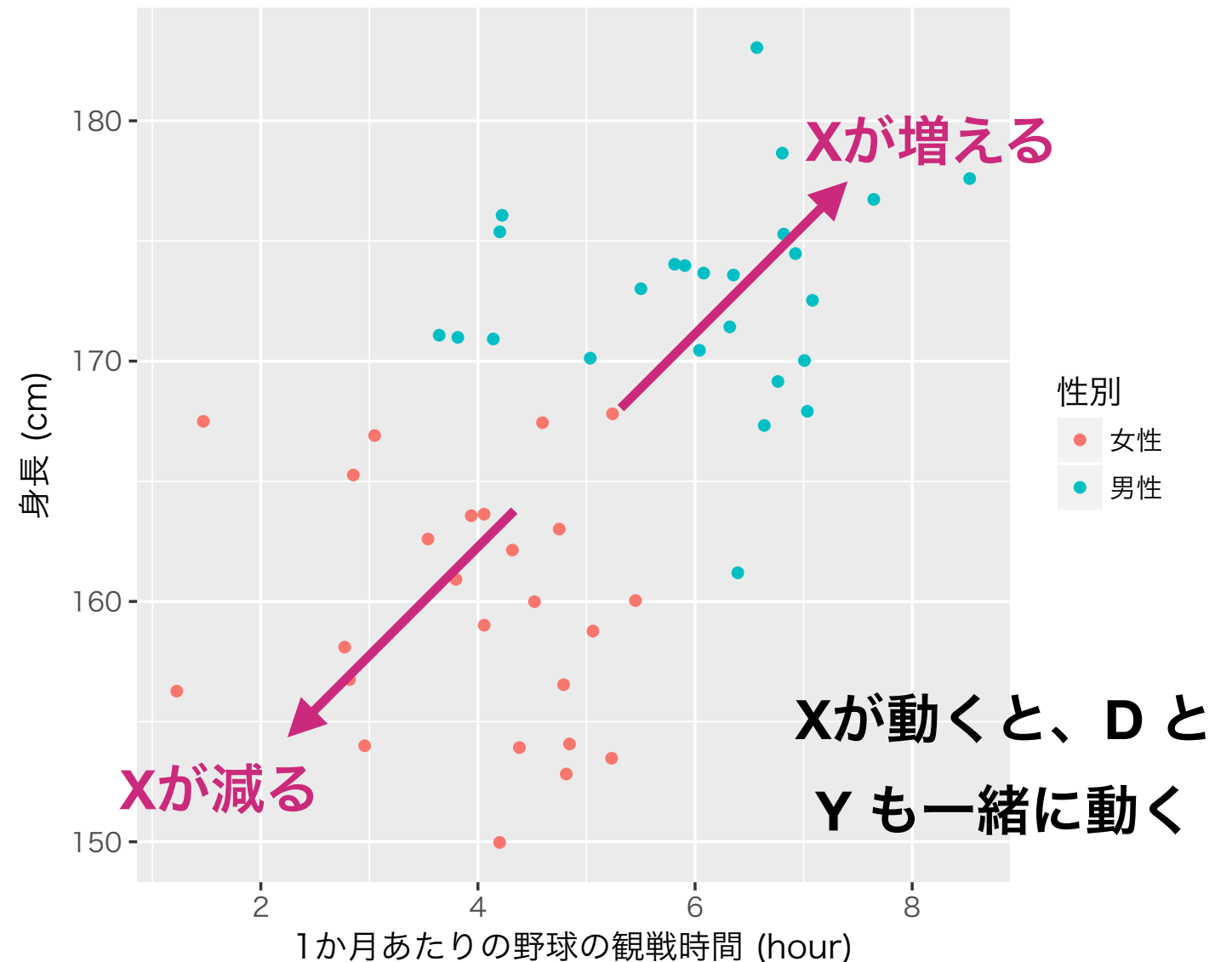
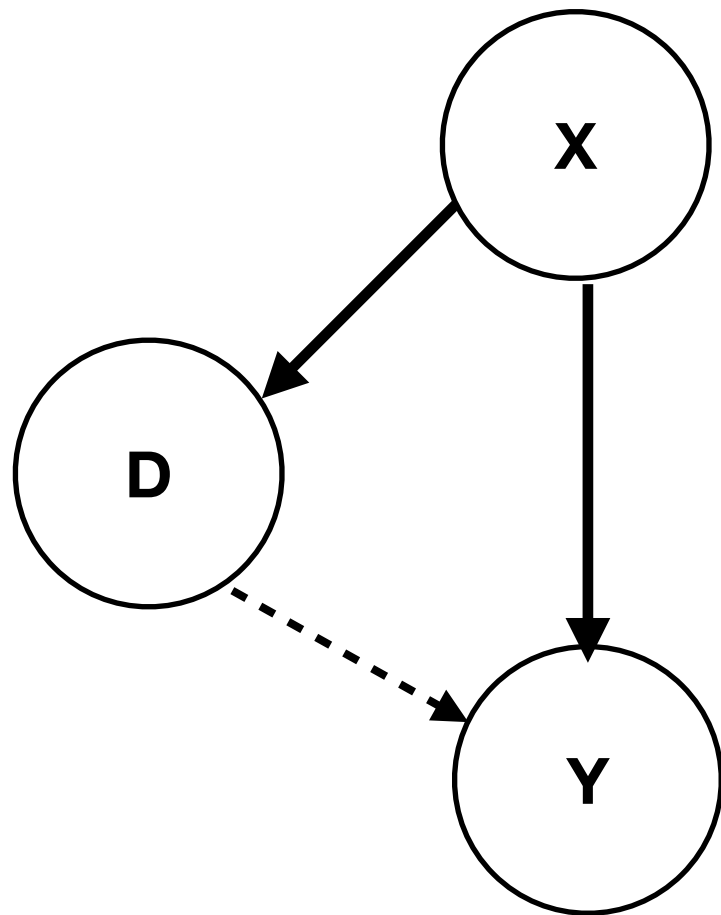
合流点 X



媒介変数 X

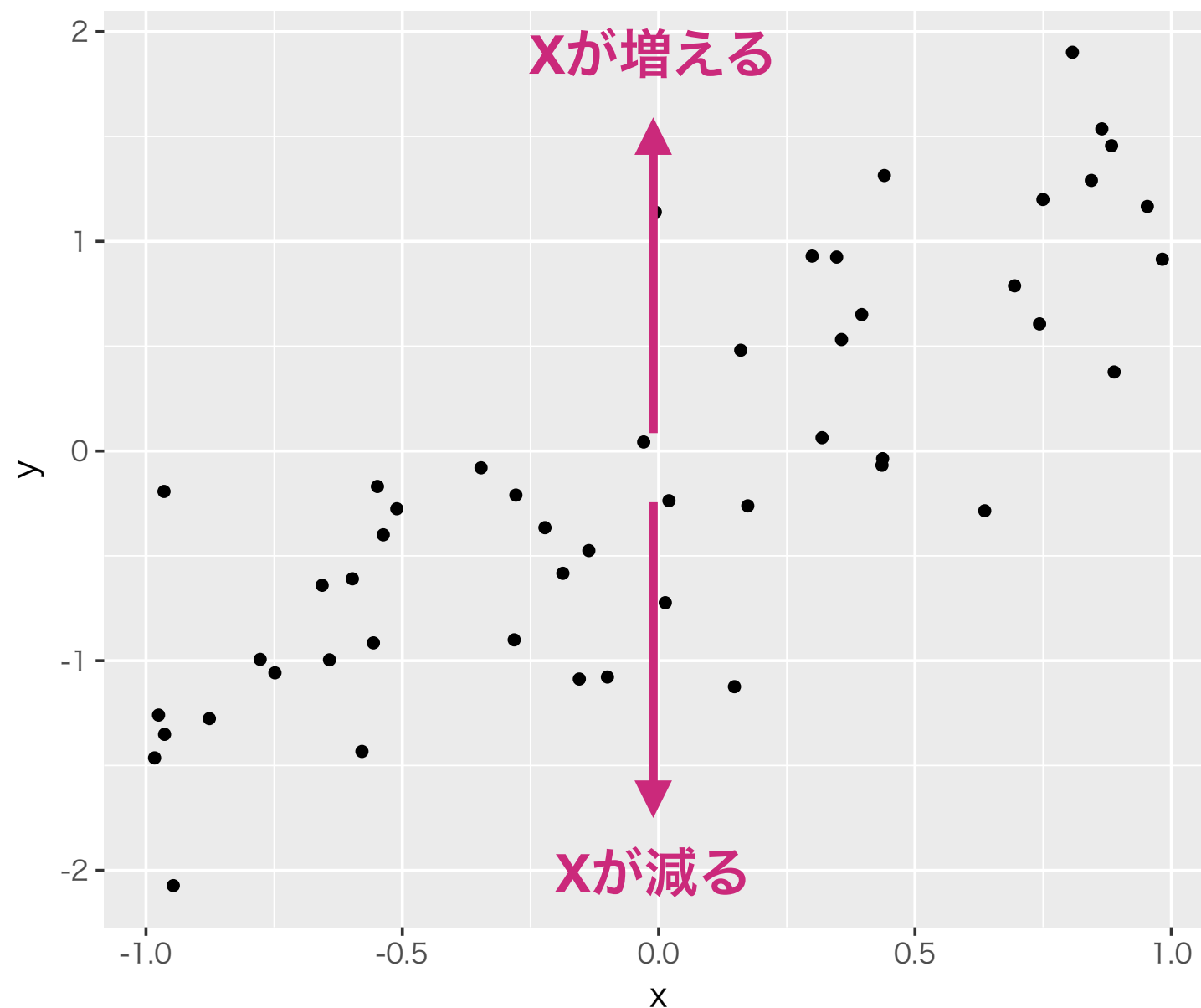
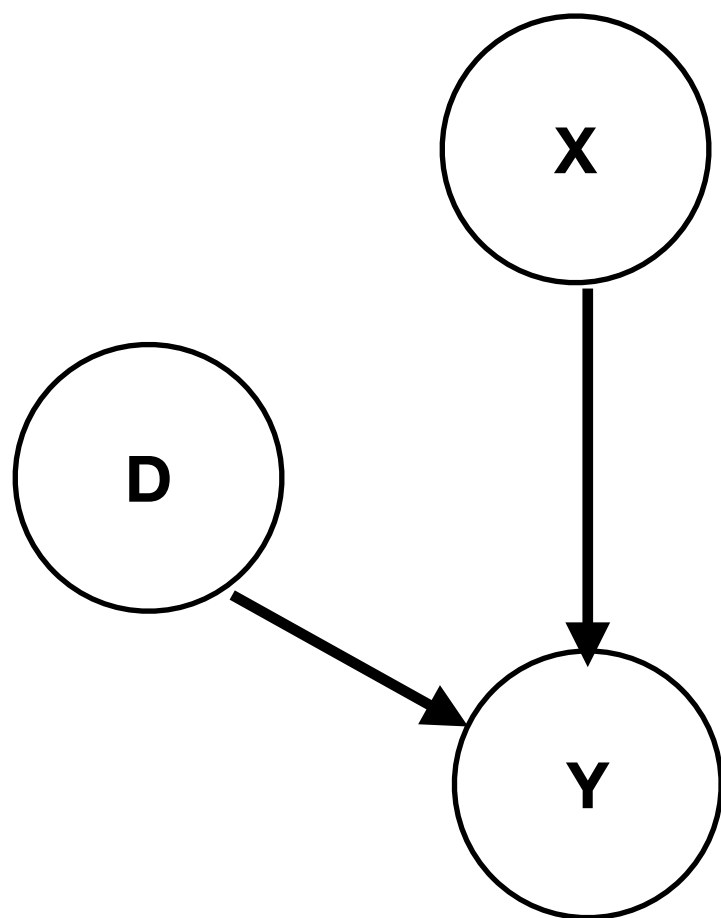


Xが交絡変数のとき



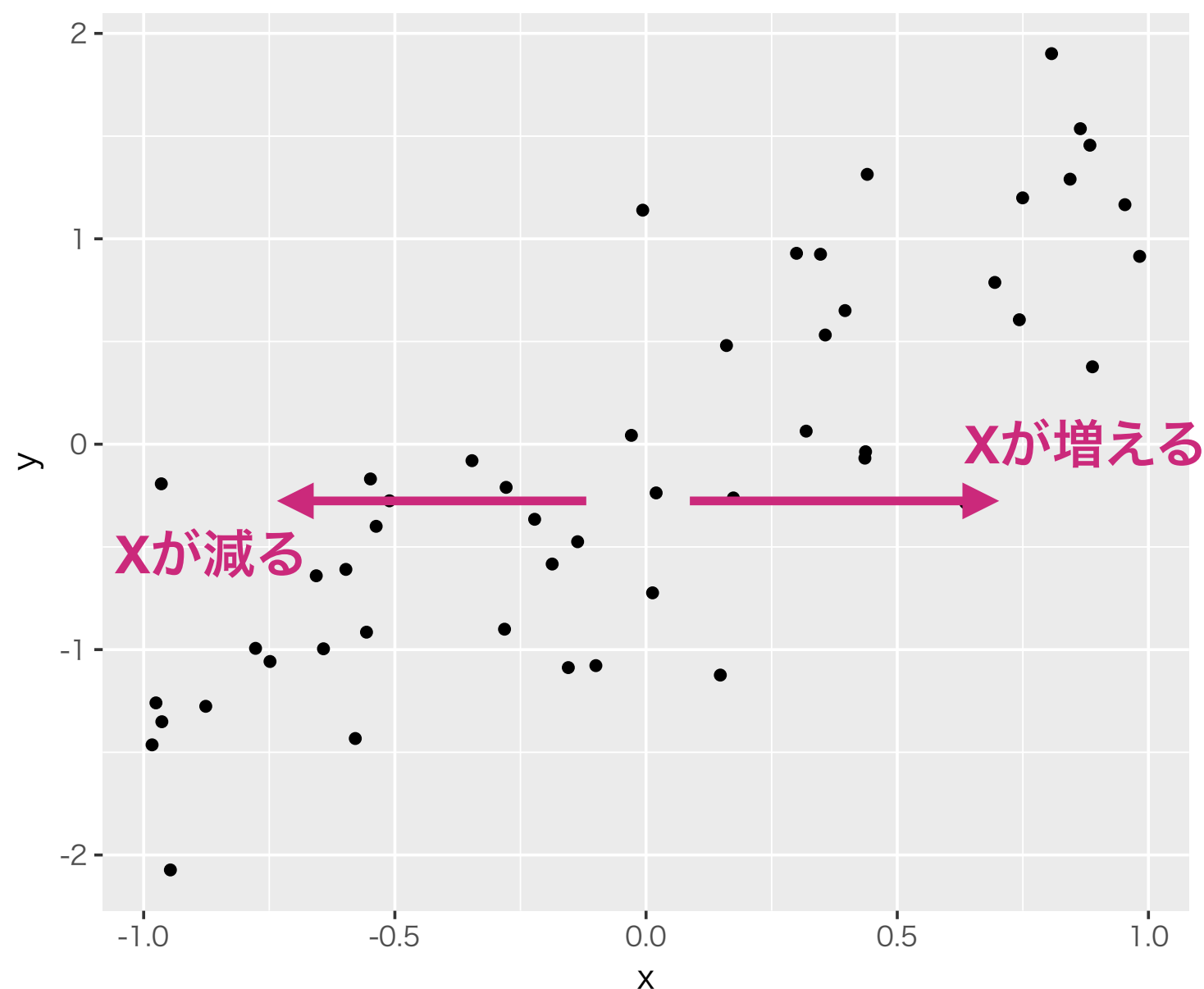
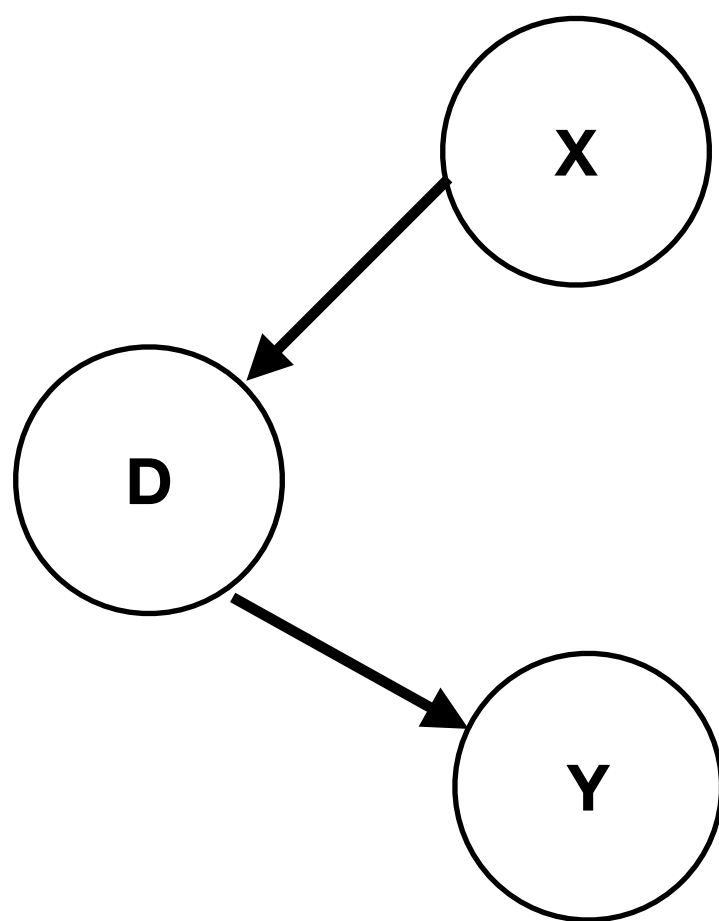
- バックドアが開いていると、Zの変化がXとYの変化を同時に引き起こす
- Y を Xだけに回帰すると、バイアスが生じる

Xが交絡ではない場合 (1)



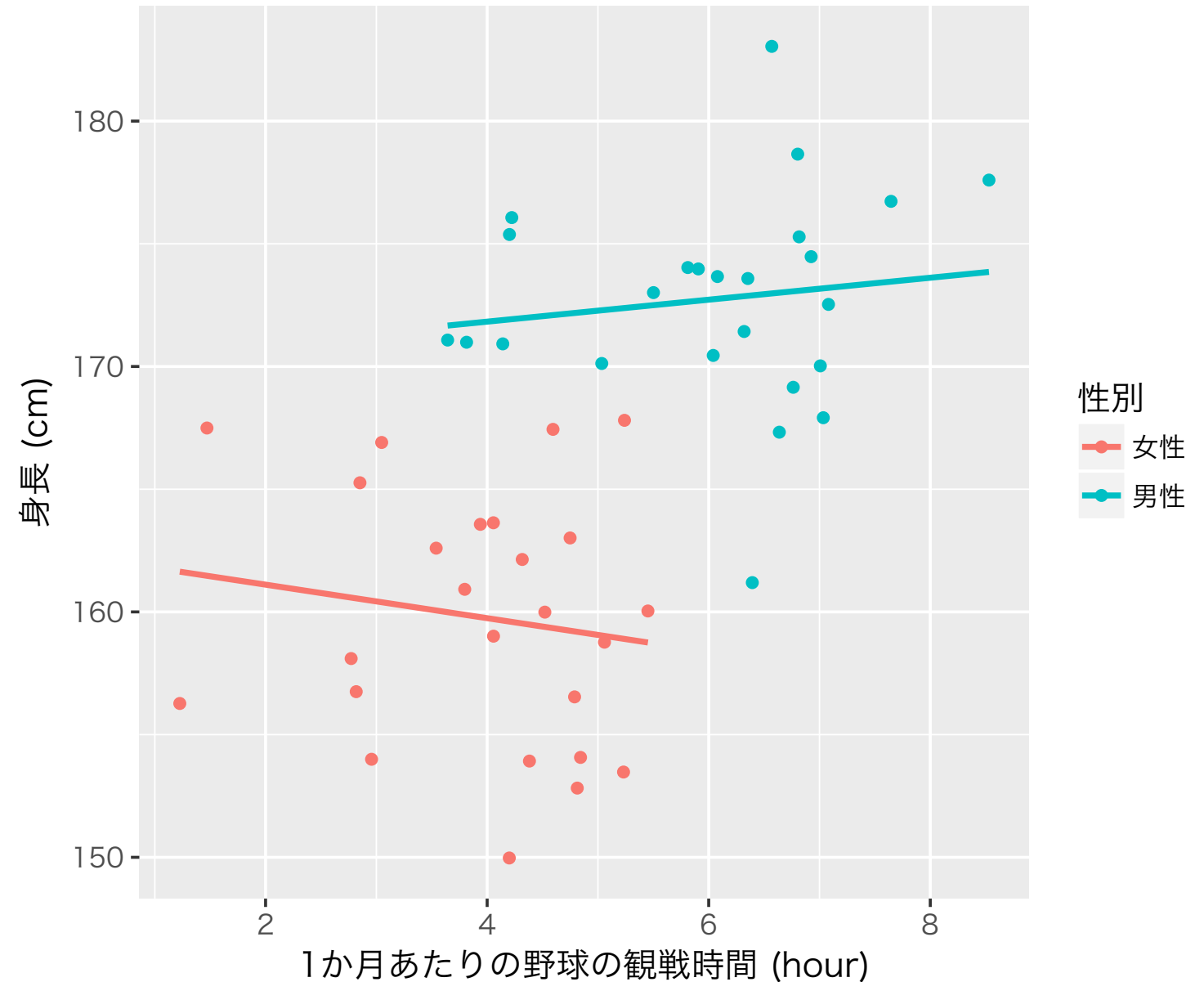
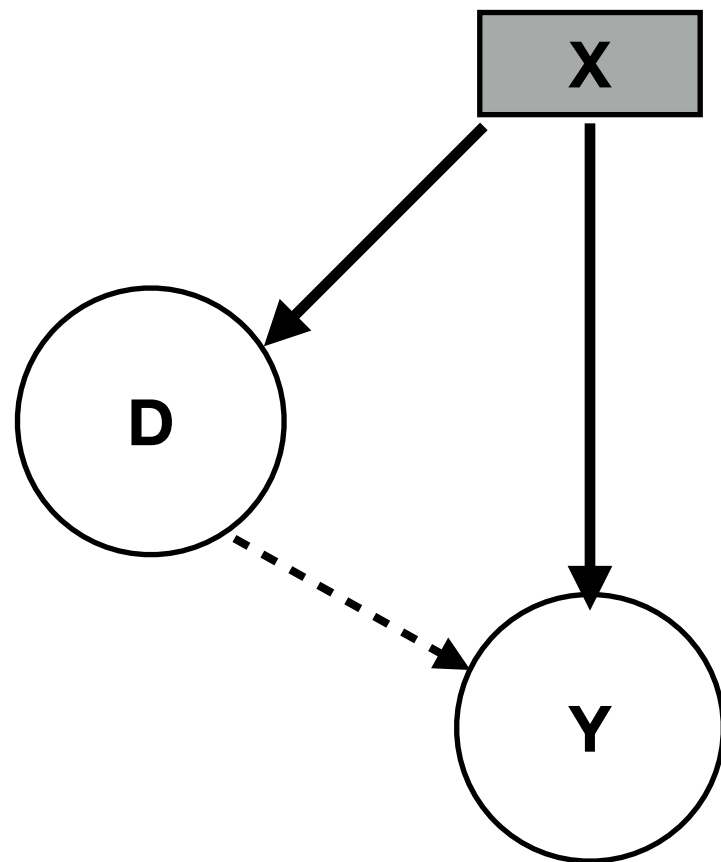
- Xの変化は、Dの変化には影響しない

X が交絡ではない場合 (2)



- Xの変化は、Yの変化には影響しない

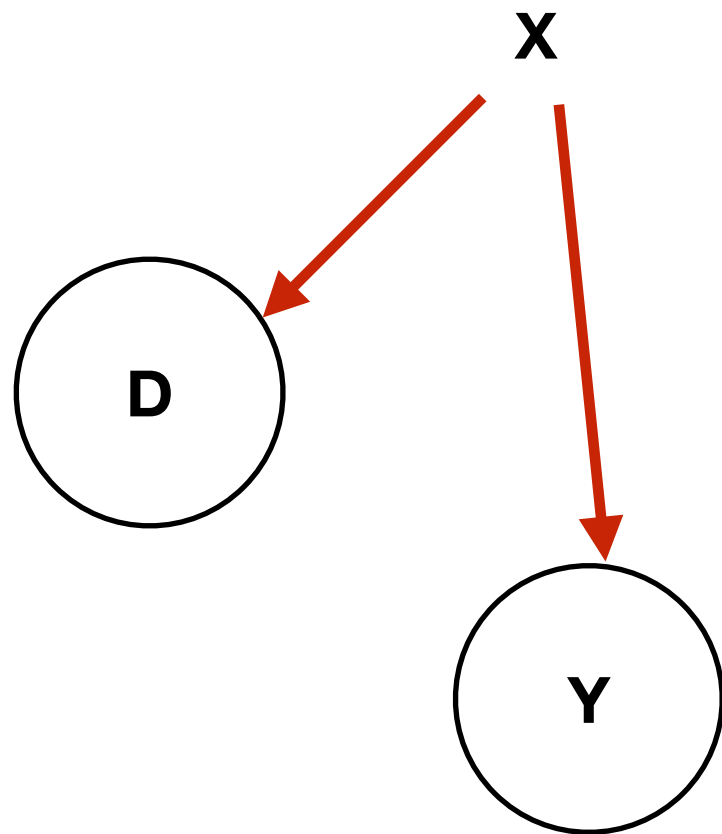
バイアスを取り除くには？



- Xの値を「固定」すればよい
 - ▶ Xをコントロールした重回帰分析

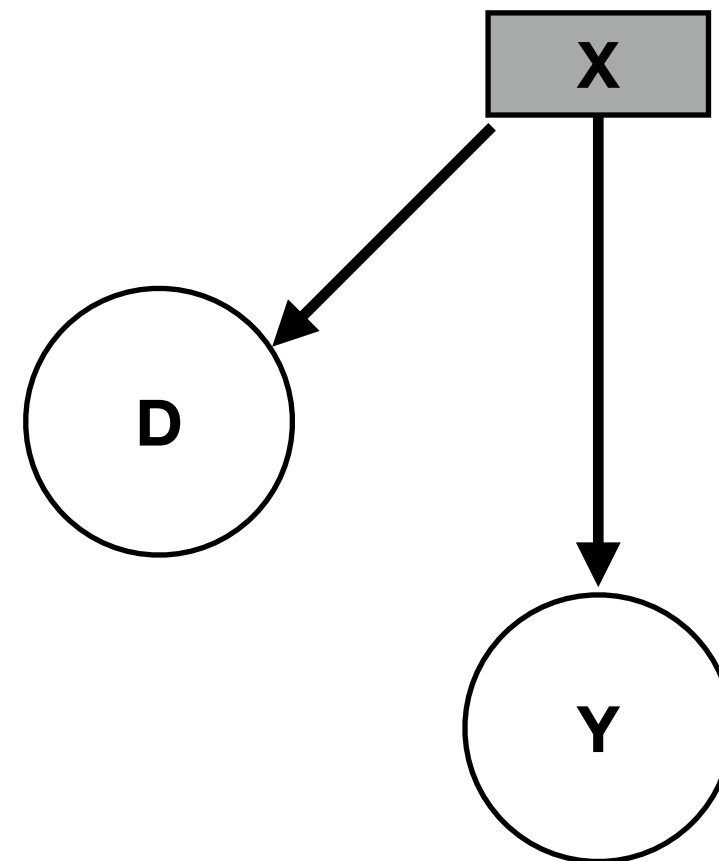
バックドアを閉じる

Xなしの回帰



バックドアが開いている：
X が考慮されていないので、バックドア
を通じたXの影響をDの影響だと見誤る

Xを含む回帰

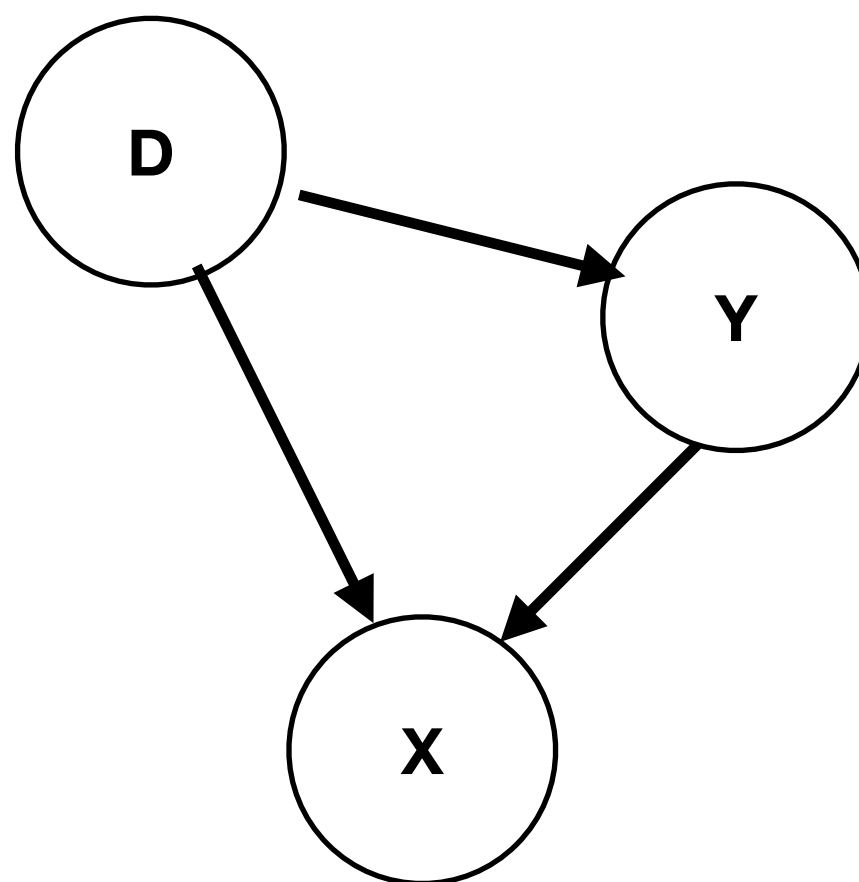


バックドアが閉じて（塞がれて）いる：
X が考慮されているので、バックドア経
路はDの影響と見なされない

回帰分析における交絡変数の扱い方

- 交絡はコントロールせよ！
 - ▶ 交絡をコントロールすれば、セレクションバイアスは除去できる
 - ▶ 交絡をコントロールし損ねると、**脱落変数バイアス** (omitted variable bias; OVB) が生じる

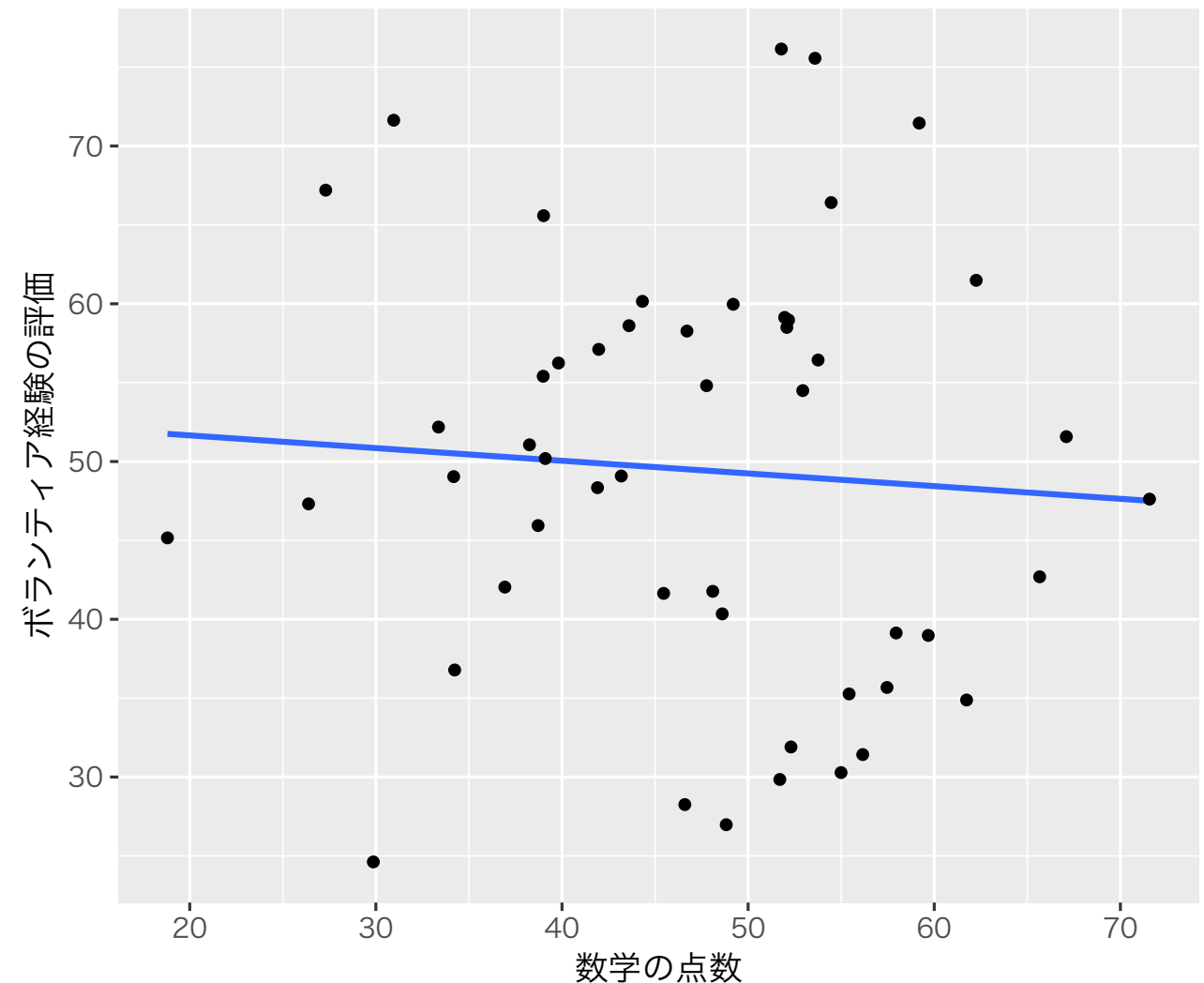
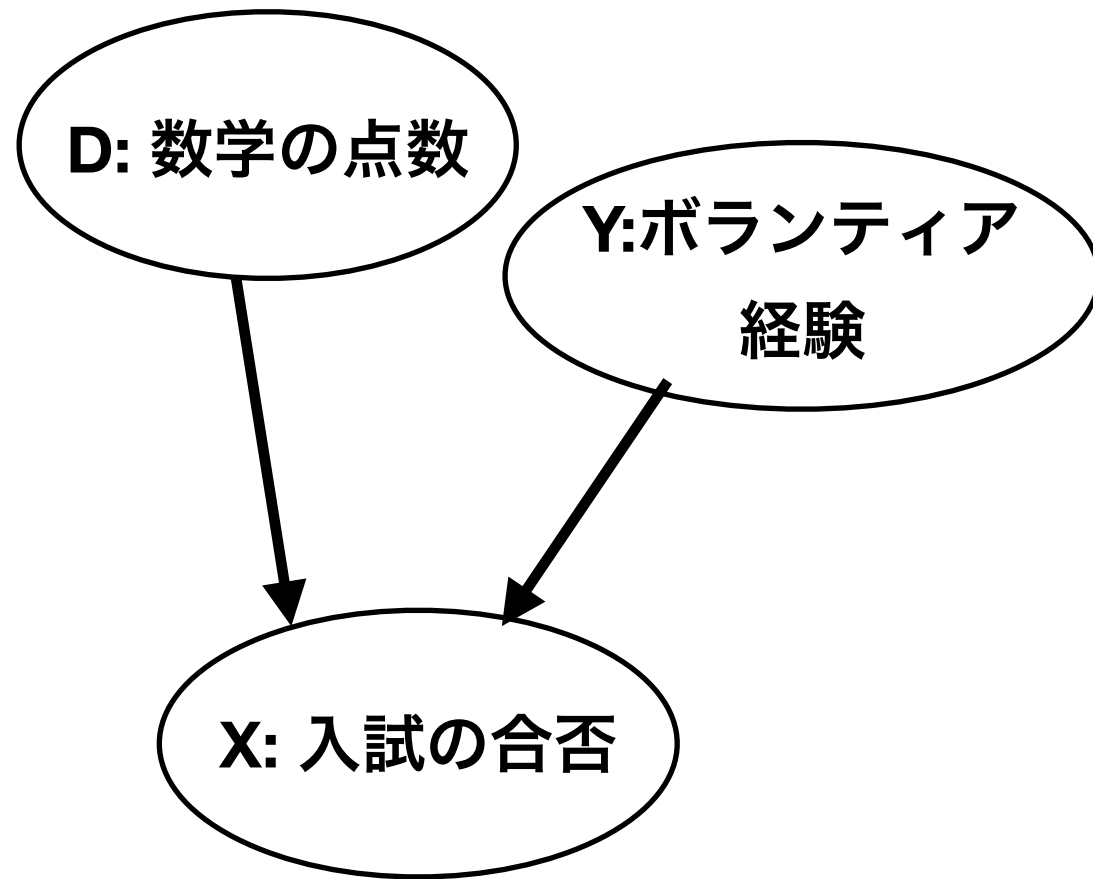
Xが合流点のとき



- Xを無視した単回帰で、DのYに対する因果効果を推定できる

合流点を統制すると何が起こる？ (1)

例：アメリカ合衆国の大学入試

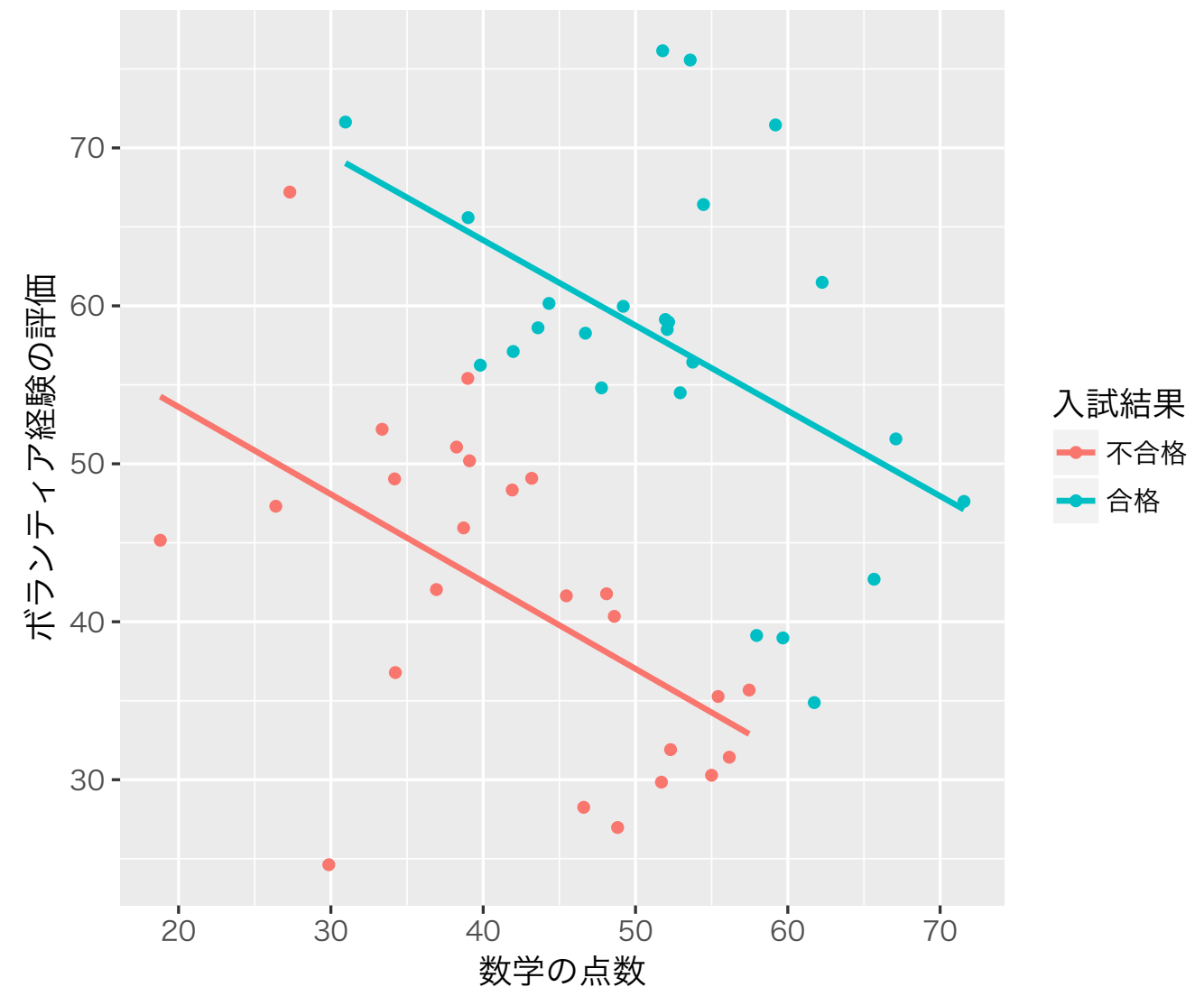
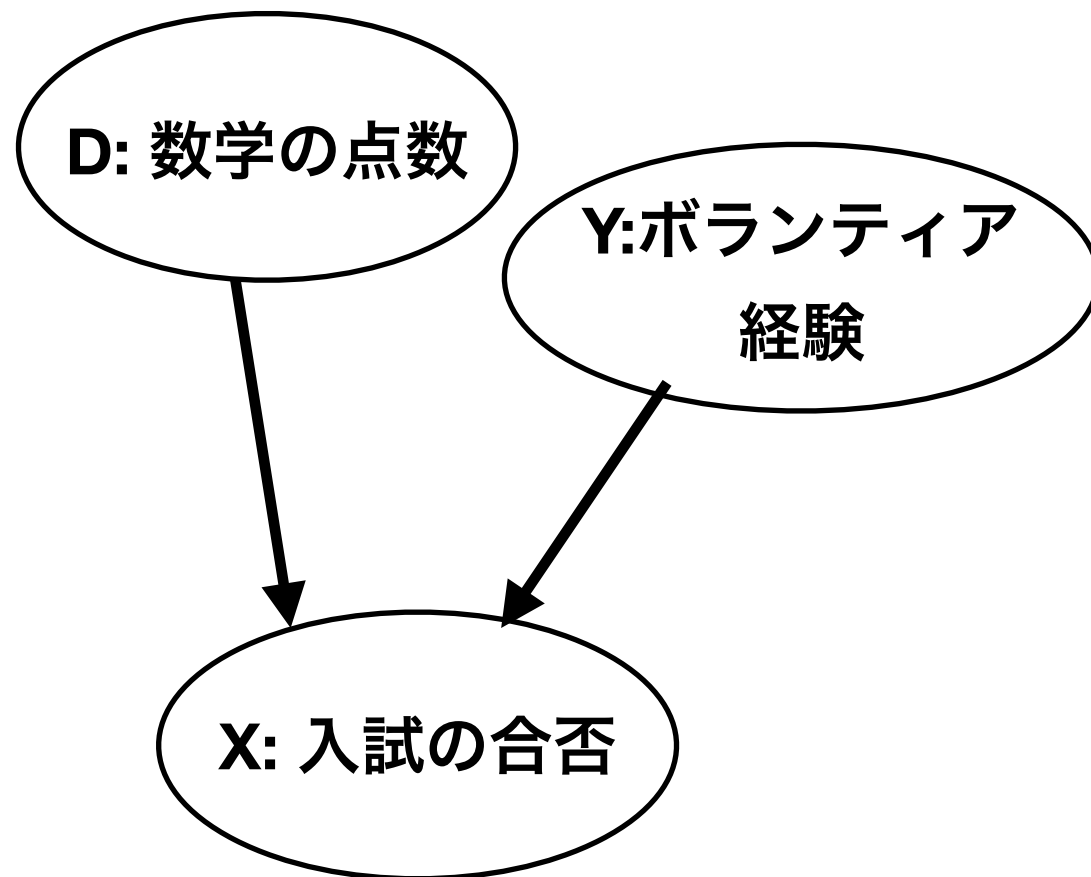


- 入試の合否は、数学の点数とボランティア経験の評価によって決まる
(架空のデータ)

▶ D から Y への因果効果はない

合流点を統制すると何が起こる？ (2)

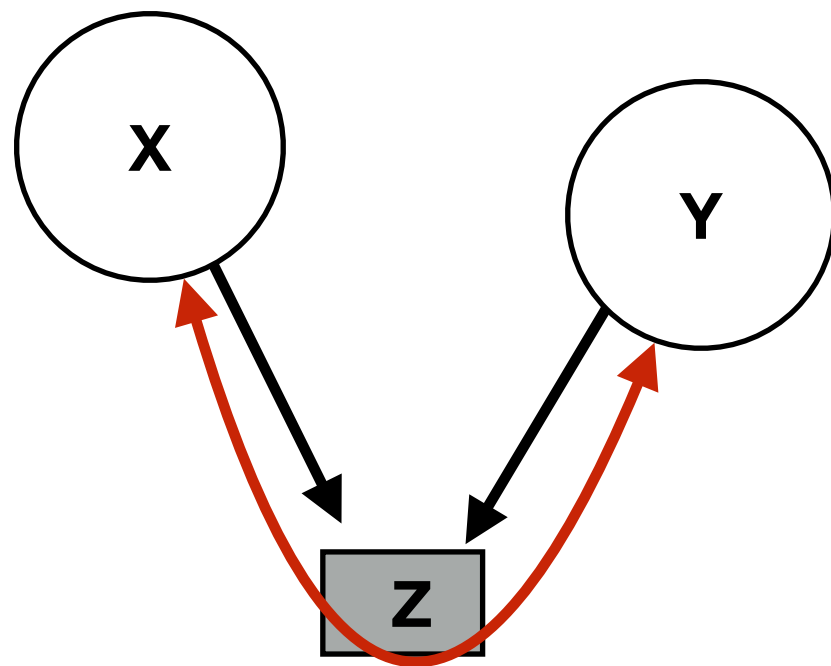
例：アメリカ合衆国の大学入試



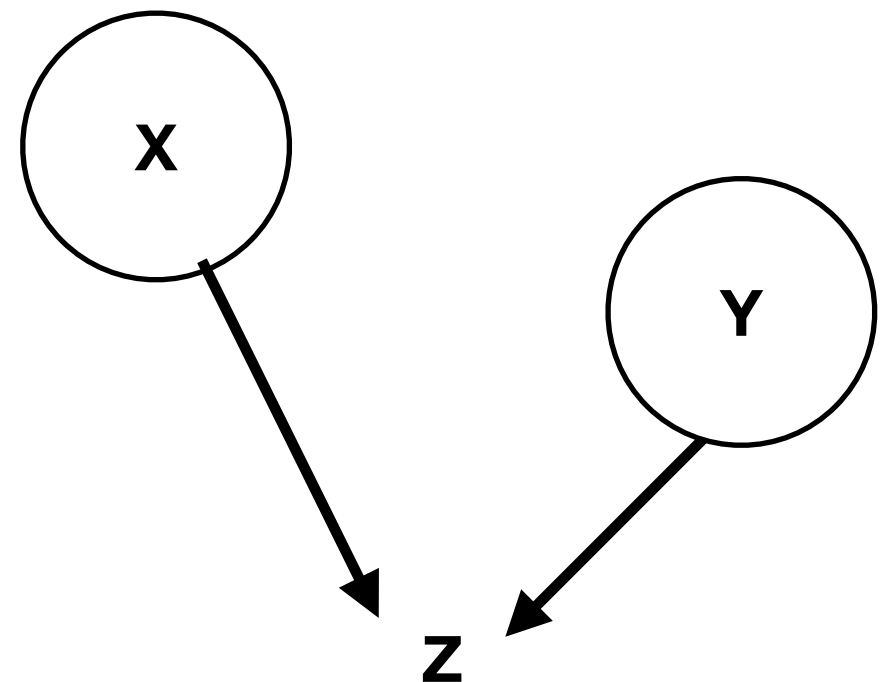
- 合流点Xを統制すると、重回帰で因果効果ではない効果を捉えてしまう

合流点とバックドア経路

Xを含む回帰



Zを含まない回帰



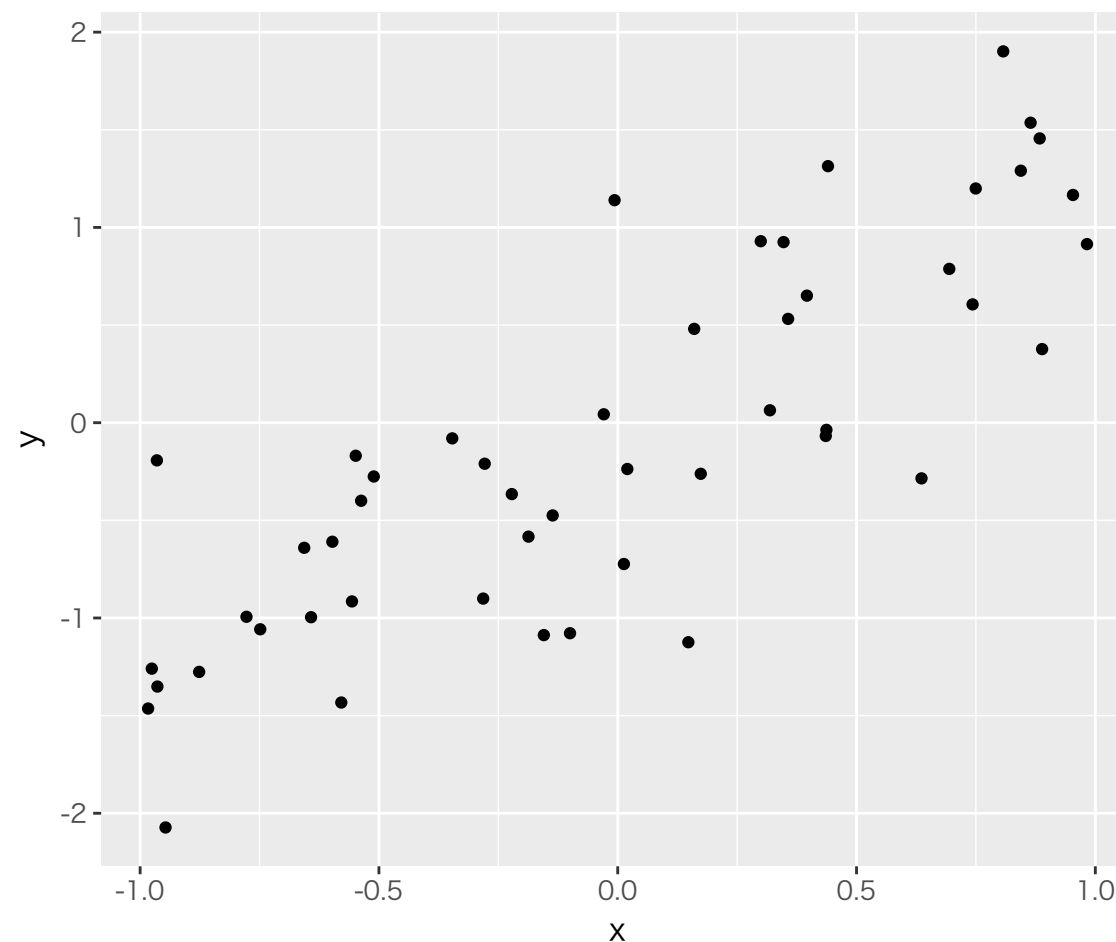
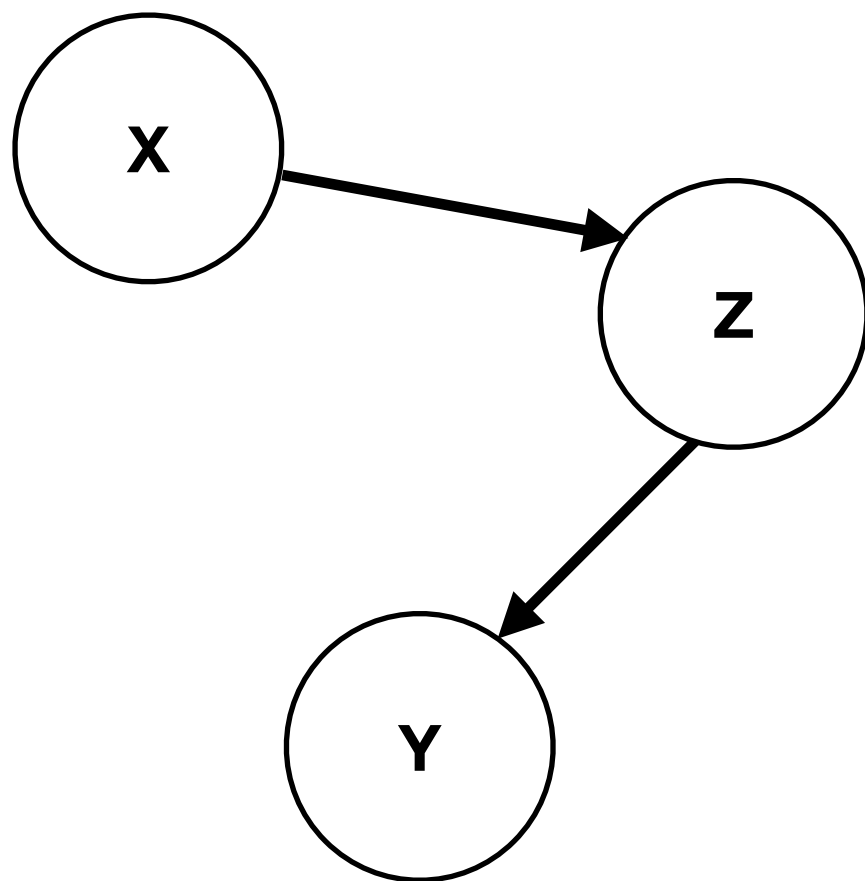
バックドアが「開いて」しまう：
DとYに関係はないのに、経路が繋がってしまう

バックドアは存在しない

回帰分析における合流点の扱い方

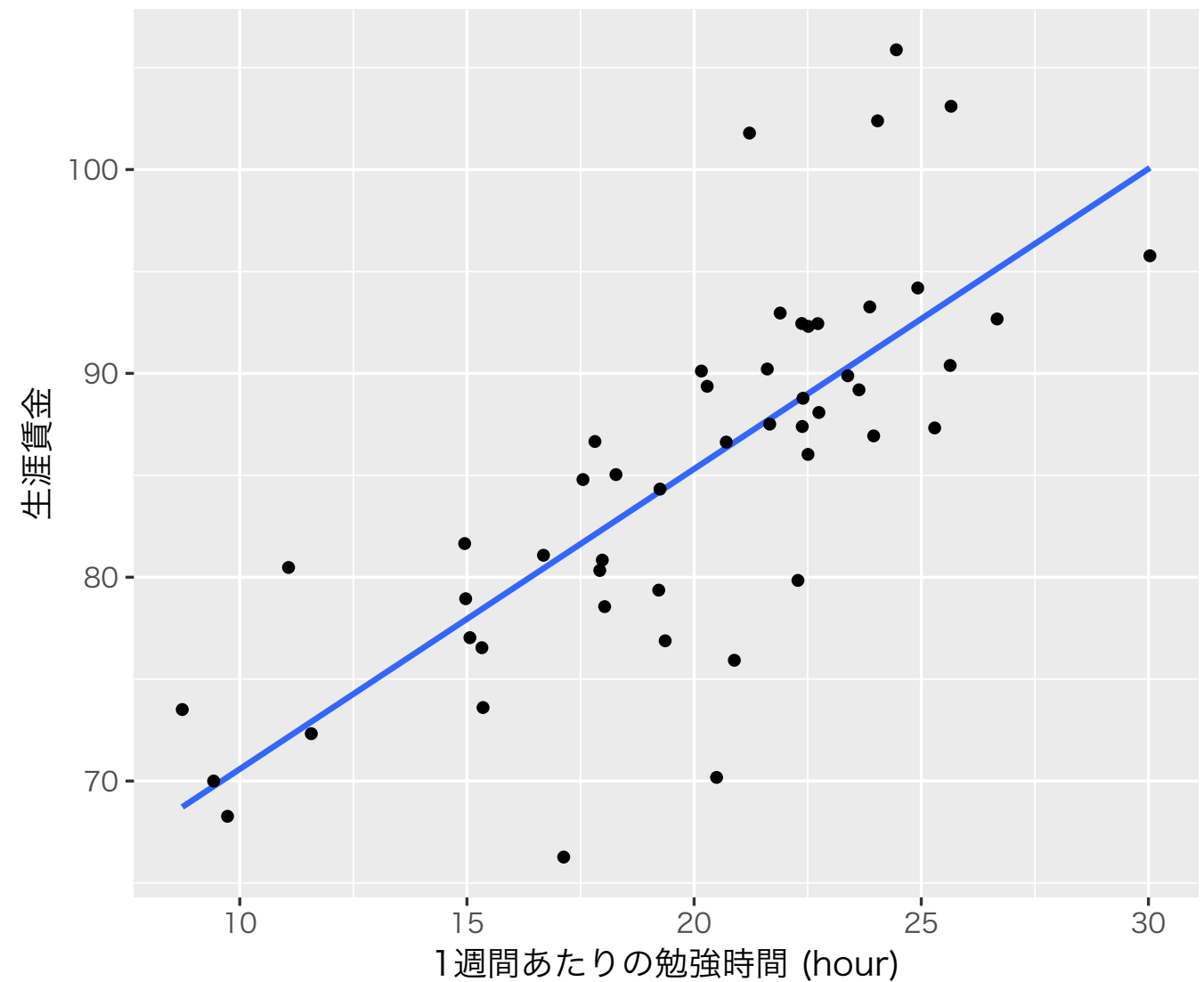
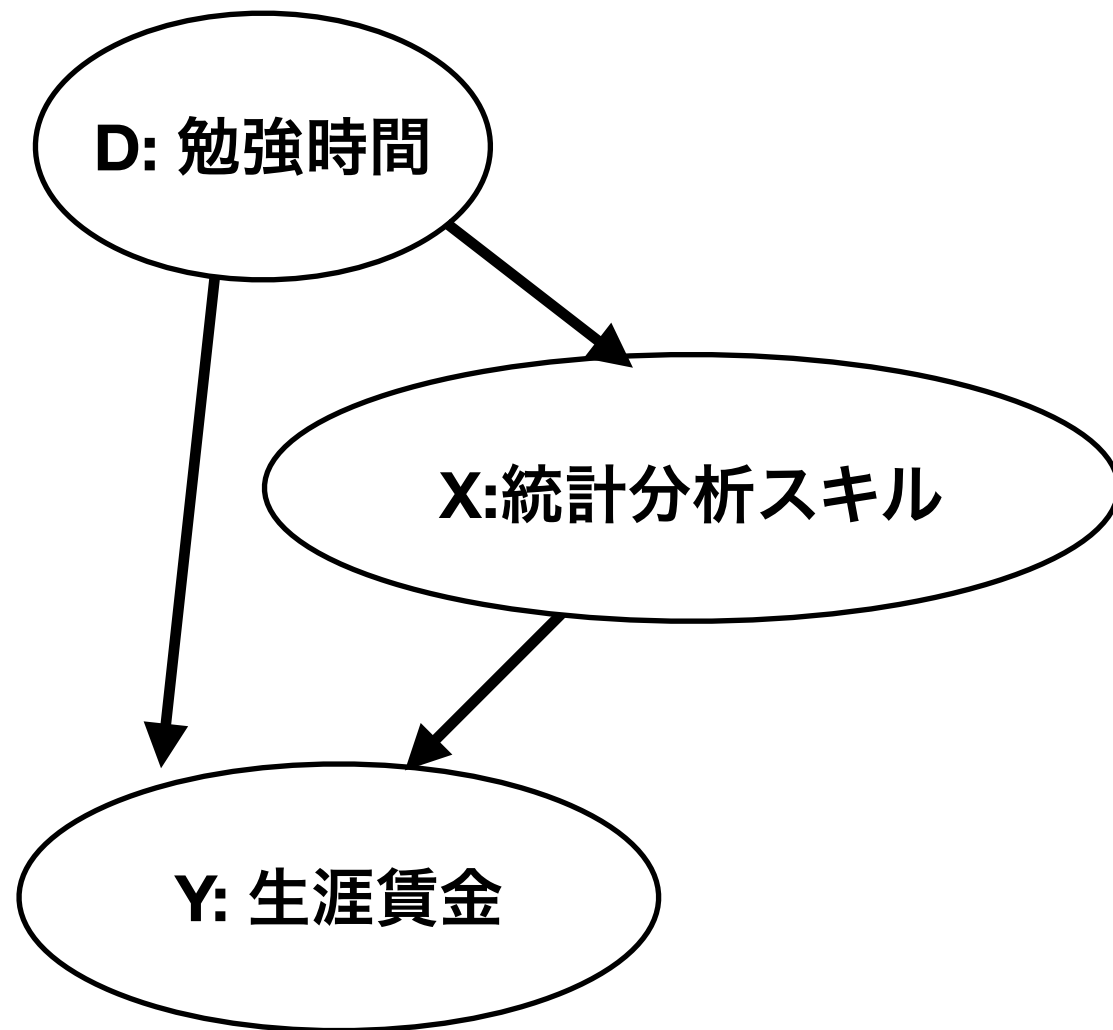
- **理論的に**考えて合流点だと思われる変数は、**回帰分析から外す**

Zが媒介変数のとき

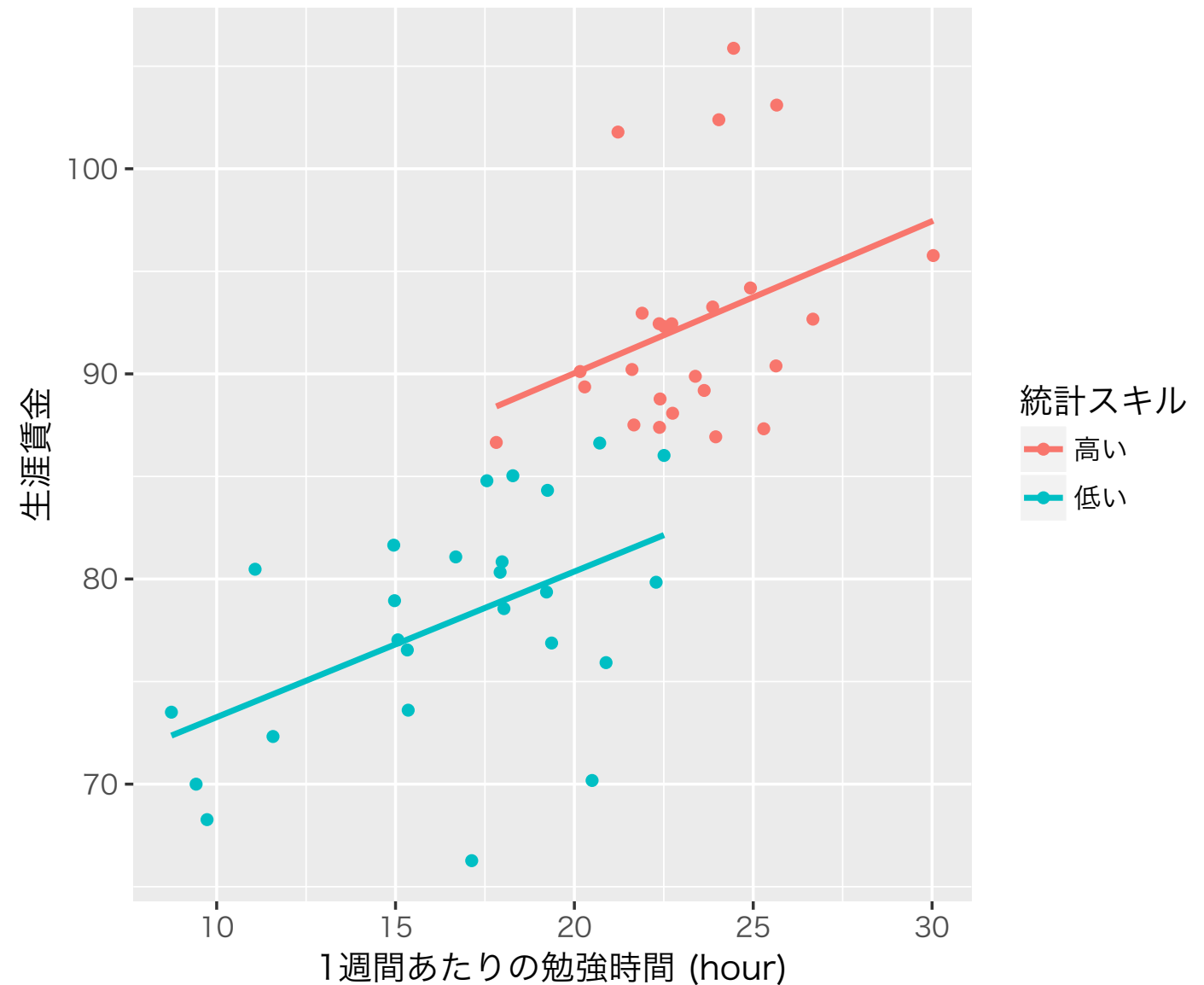
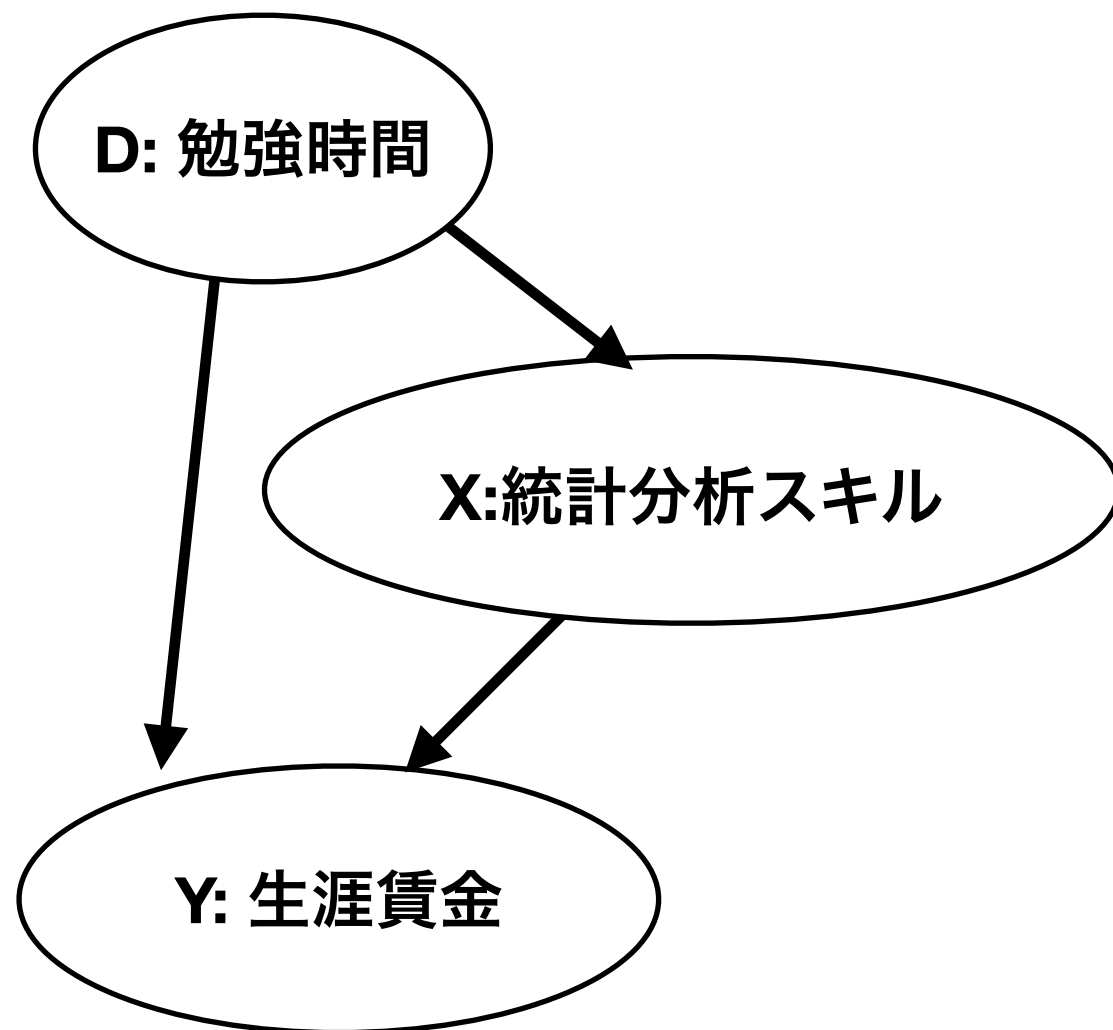


- Zを含まない単回帰モデルで、因果効果を推定できる

媒介変数を統制すると何が起こる？ (1)



媒介変数を統制すると何が起こる？ (2)



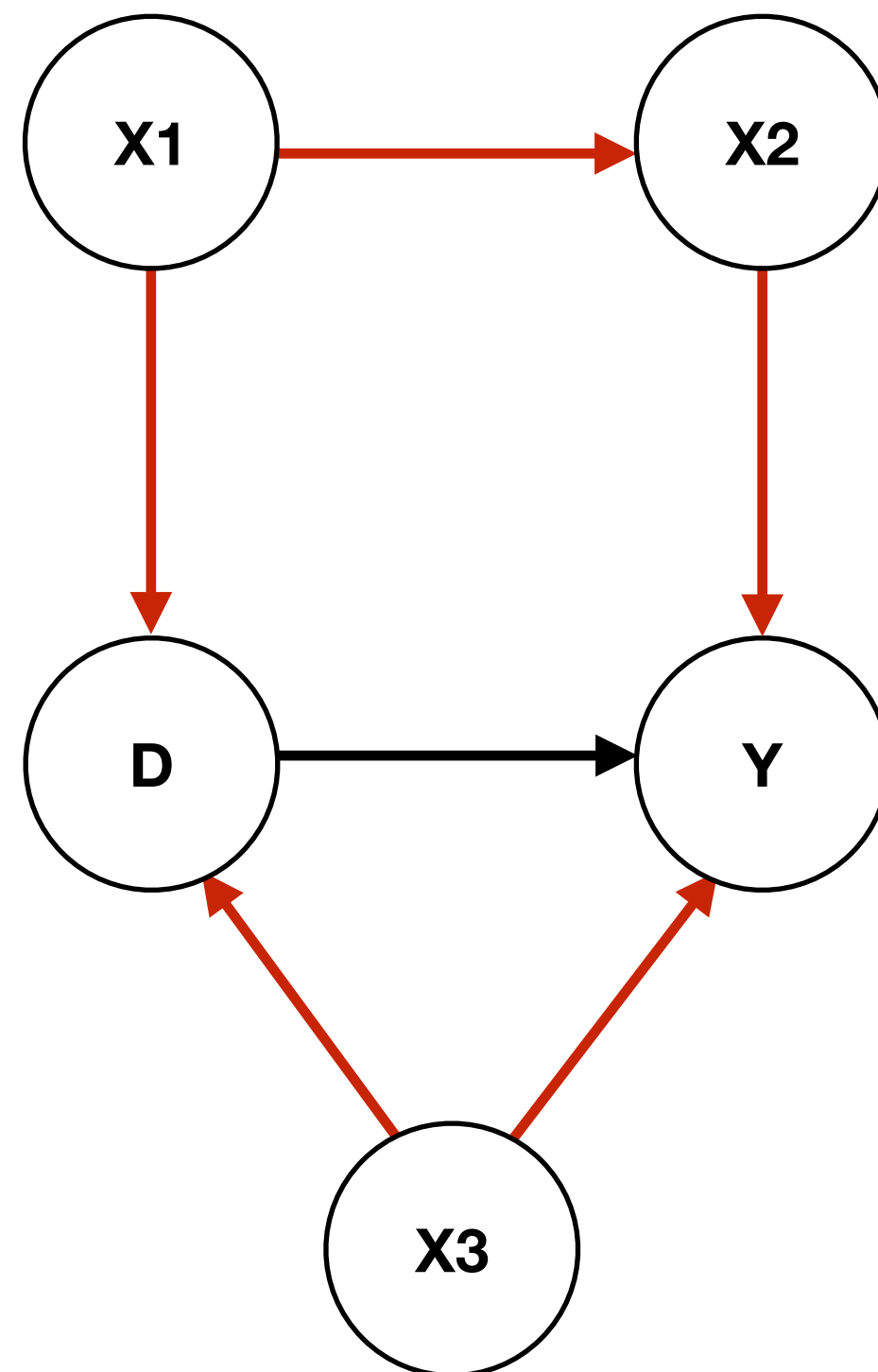
- 媒介変数 X を統制すると、 D から Y の経路の一部が塞がれてしまう
 - ▶ 因果効果が過小評価される：処置後変数バイアス

回帰分析における媒介変数の扱い方

- **理論的に**考えて媒介変数（中間因子）だと思われる変数は、**回帰分析から外す**

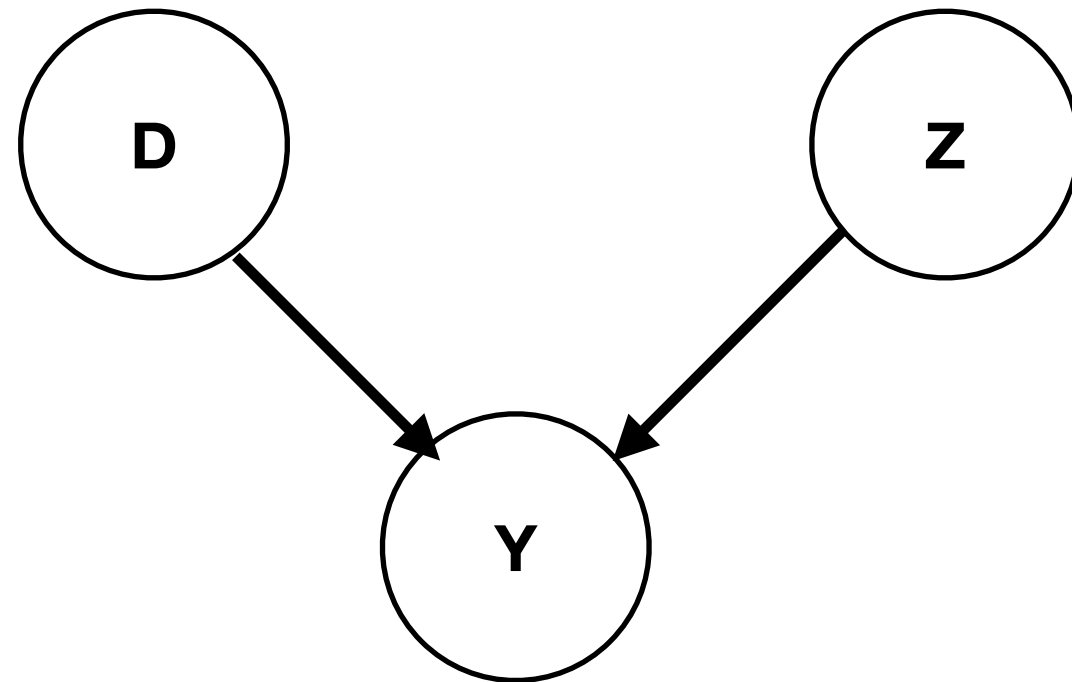
変数の数が多いとき

- 右の図のバックドア経路
 - ▶ $D \leftarrow X1 \rightarrow X2 \rightarrow Y$
 - ▶ $D \leftarrow X3 \rightarrow Y$
- バックドア経路をすべて閉じればよい
 - ▶ $X1$ と $X3$ を統制する
 - ▶ $X2$ と $X3$ を統制する
 - ▶ $X1$ と $X2$ と $X3$ を統制する



その他の場合は？

- 交絡でもなく、合流点でもなく、媒介変数でもないZを統制すると何が起こる？
- 推定の効率性が落ちる（標準誤差が大きくなる）が、推定にバイアスは生じない



因果推論における回帰分析

- 回帰分析は、統計的因果推論における基本ツール
 - ▶ RCT でも使える
 - ▶ 重回帰分析でセレクションバイアスを除去できる（こともある）
 - 処置後変数バイアス（媒介変数、合流点の誤投入）に注意
 - ▶ この授業でこれから学ぶ手法は、回帰分析の応用

因果推論における回帰分析の問題点

- 「コントロール」によってセレクションバイアスを取り除けるとは限らない
 - ▶ 交絡因子を誤解している
 - 交絡を交絡ではないと判断：脱落変数バイアス
 - 処置後変数を交絡だと判断：処置後変数バイアス
 - ▶ 交絡が未観測・観測不能
 - どうすればいいの？

次回予告

Topic 5. 傾向スコア