計量経済学応用

11. 回帰分析におけるバイアス

矢内　勇生

2018年5月21日

高知工科大学　経済・マネジメント学群

今日の目標

- 重回帰分析で何をコントロール（統制）すべきか理解する

- 重回帰分析で生じ得るバイアスについて理解する

統計分析の目的は？

- 実験：因果関係を調べるための最善策
- しかし、どんな問題でも実験できるわけではない
- 調査・観察・観測データ（observational data）に頼るしかない

統計分析

- 比較的大きな標本サイズのデータを使って理論を検証する
- 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：重回帰分析
  - パネルデータ分析

何を検証する？(1)

- 検証したい理論：　「XがYを引き起こす」
- この関係は決定的 (deterministic) か？
- 例：「教育が政治参加を促す」
- 変数の操作化
  - 教育：大卒か否か
  - 政治参加：国政選挙での投票参加

何を検証する？(2)

- 決定論的理論: 「大学の学位は国政選挙で投票するための必要十分条件である」

何を検証する？(2)

- 決定論的理論: 「大学の学位は国政選挙で投票するための必要十分条件である」
- 大卒だが投票していない人を「1人だけ」見つけたらどうする？

何を検証する？(2)

- 決定論的理論: 「大学の学位は国政選挙で投票するための必要十分条件である」

- 大卒だが投票していない人を「1人だけ」見つけたらどうする？

- 決定論では、理論を否定する必要がある

何を検証する？(2)

- 決定論的理論: 「大学の学位は国政選挙で投票するための必要十分条件である」

- 大卒だが投票していない人を「1人だけ」見つけたらどうする？

- 決定論では、理論を否定する必要がある

- それでいい？

何を検証する？(3)

- ほとんどの場合、私たちが検証したいのは、確率論的理論：
  「大学に行くと、国政選挙での投票確率が上がる傾向にある」

- 理論的予測に合致しない人を少数見つけても、理論の否定にはならない

  - 十分大きな標本サイズで、少人数が理論に合致しなくても、大きな傾向に影響はない

- 大卒と大卒未満の2つのグループで、平均すると大卒の方が投票率が高いことを示せればよい

条件付け：「他の条件が等しければ」

- 2つ（以上）のグループを比較する
- 社会科学では、以下のような異質な個体を比較する
  - 人間
  - 国家
- 通常、調べている要因以外の「他の条件」は等しくない！
  - 大卒と大卒未満では、「親の年収」に違いがあるかもしれない
- 「他の条件を等しい(ceteris paribus)」状況で比較したい

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」
- 結果変数： 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）

コントロール・条件付け

- 変数 Z を統制（コントロールする）：Zは統制（コントロール）変数と呼ばれることも

- Zは複数あってもよい: Z1, Z2, …

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

- 複数の要因を統制するためには、大きな標本サイズが必要

  - Ｎ＝２で一人は女性、もう一人は男性のとき、性別を統制できる？

どの変数を統制する？

- 重回帰で使う変数は何？
  - 結果変数（理論における結果）：絶対に必要
  - 主な説明変数（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
    - 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、有向非循環グラフ）

  - 回帰分析でもこの考え方は便利

回帰分析

- 線形回帰モデル（最小二乗法で推定）を考える
  - Y: 結果変数
  - X: 主な説明変数
  - Z: 統制変数
- 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

回帰分析の例

- 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？

回帰分析の例

- 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!

回帰分析の例

- 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…

回帰分析の例

- 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）

傾きパラメタの p 値 < 0.01

身長 (cm)

180

170

160

150

2

4

6

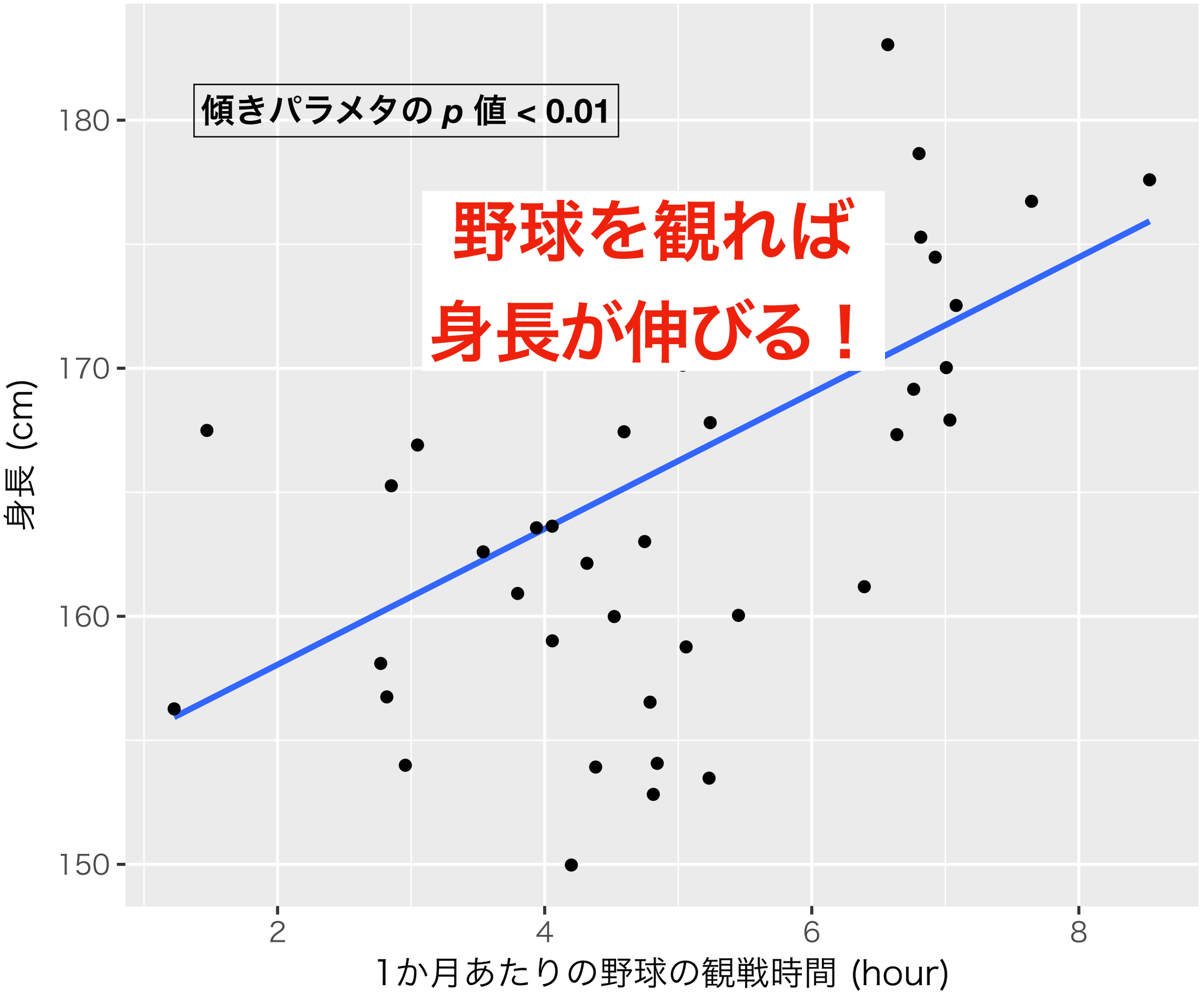
8

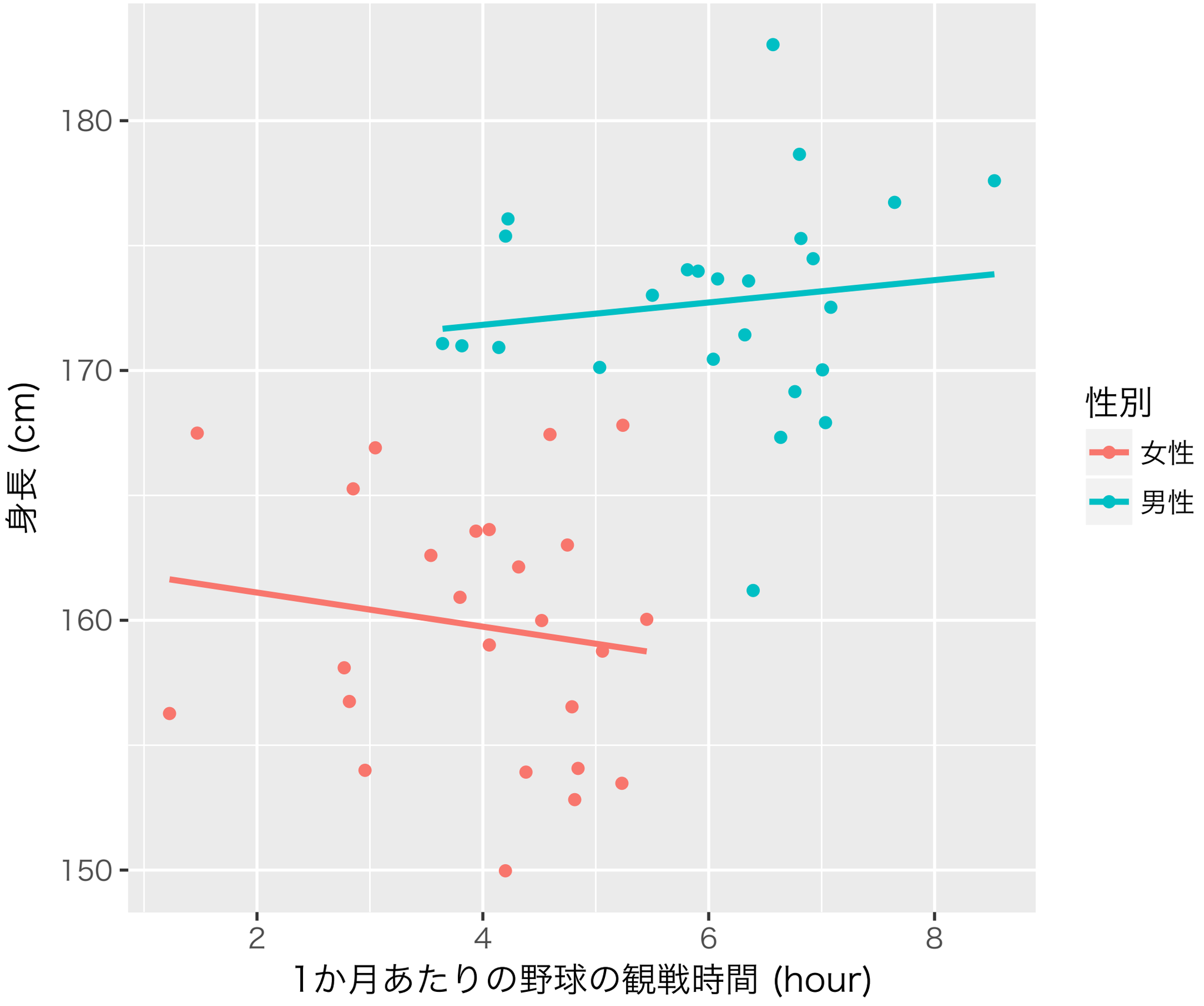
1か月あたりの野球の観戦時間 (hour)

(架空のデータ)

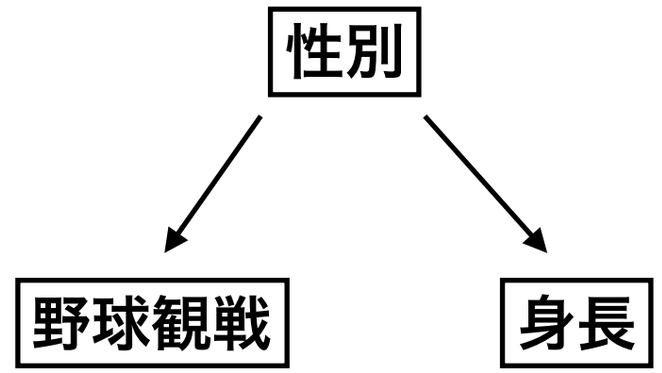
傾きパラメタの p 値 < 0.01

野球を観れば
身長が伸びる！




180
170
160
150
(cm)
2
4
6
8
1か月あたりの野球の観戦時間 (hour)
性別
女性
男性

何が問題か？



- 統制すべき「他の要因」が存在
- 女性と男性は同じではない
- 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
  - 男性の方が野球を観る
  - 男性の方が身長が高い

バックドアが開いている

- 統制変数を入れ忘れた回帰分析だと、なぜ間違えるのか？
  - 最小二乗推定量が、因果効果の推定を誤る：推定結果にバイアスが生じる
    - 欠落変数バイアス？
    - 内生性 (endogeneity)？
- 今日の話：バックドアが開いているから

バックドアと分析方法

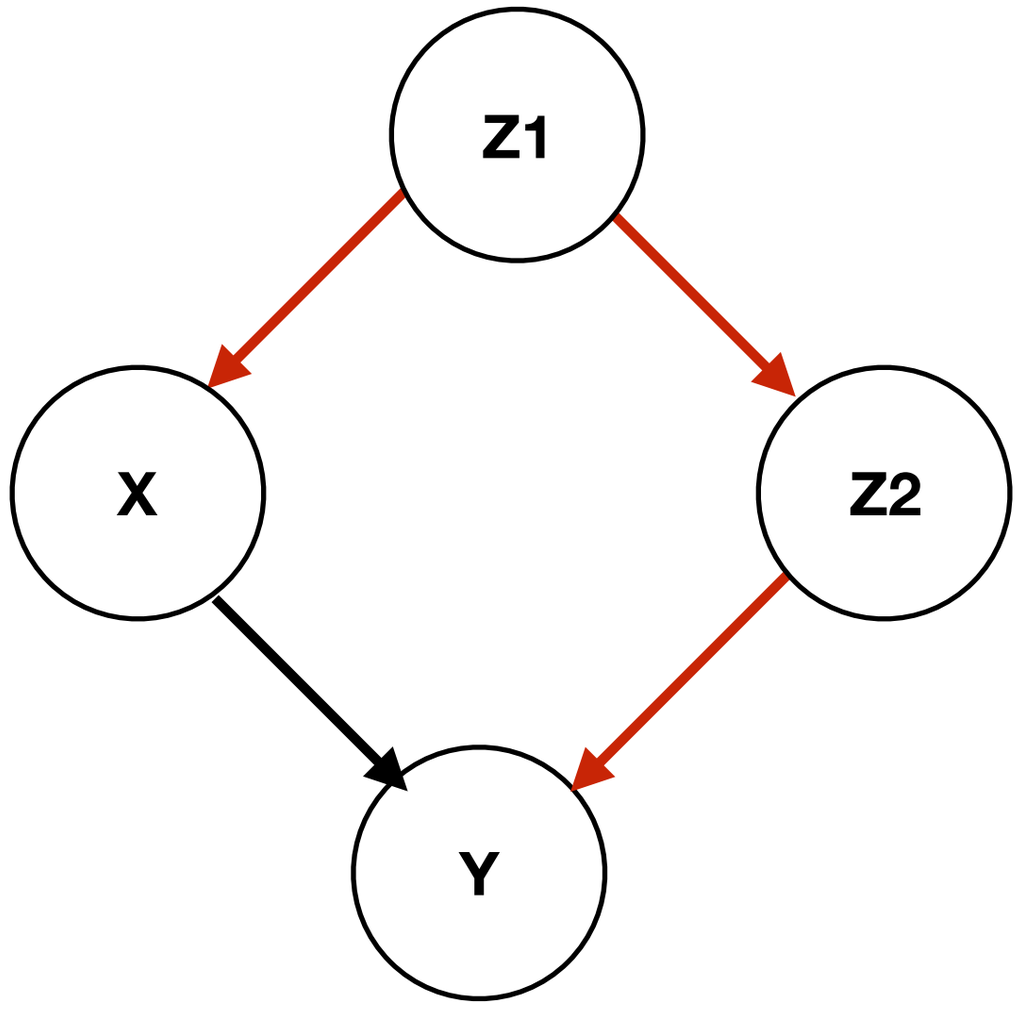
- バックドアを閉じられるとき
  - 回帰分析
  - マッチング法
  - 重み付き回帰（マッチング法と一緒に）
- バックドアが閉じられないとき
  - 操作変数法
  - その他の自然実験、準実験

バックドア基準の基礎

- X: 主な説明変数（処置変数、介入、刺激、独立変数）
- Y: 結果変数（応答変数、従属変数）
- Z: 統制変数 （コントロール、共変量 [covariates]）

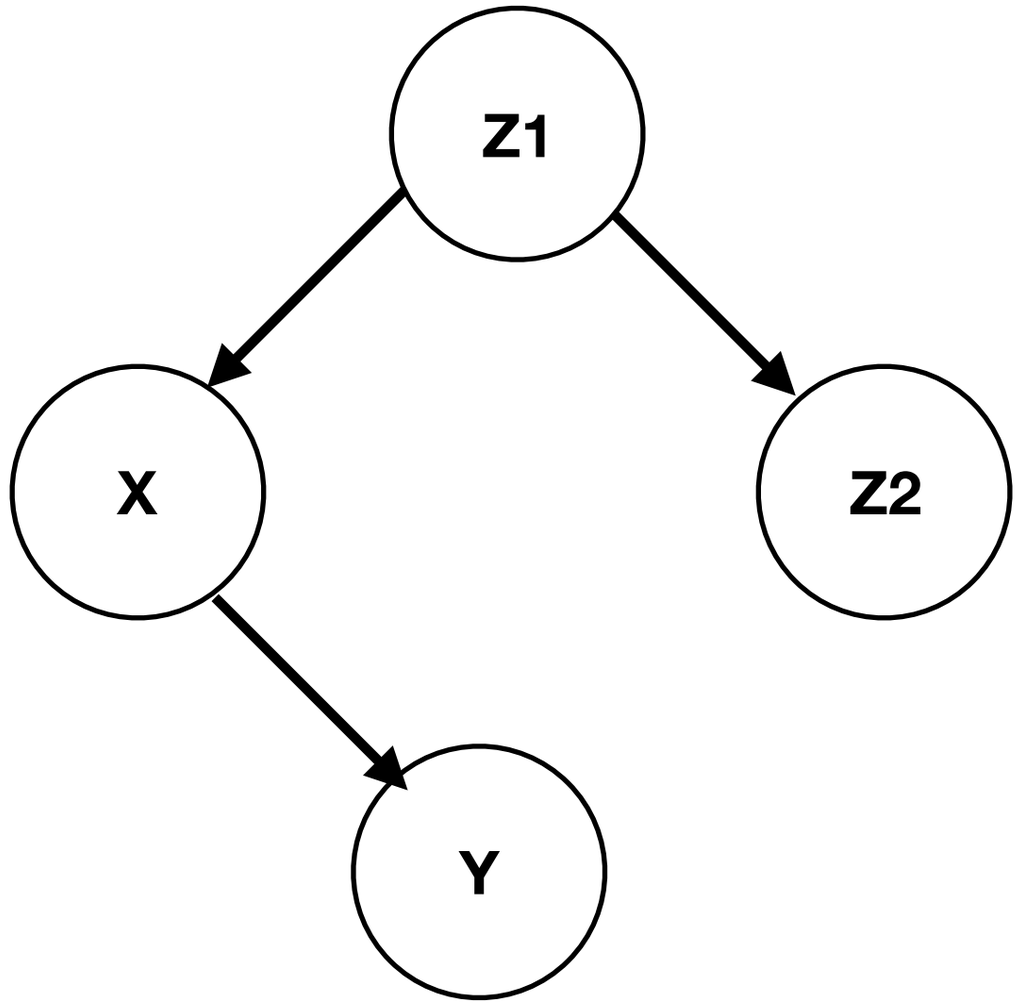
交絡変数とバックドア経路 (1)

- バックドア経路: ある変数がXとYの両者の原因となるような経路
  - X ←Z1 → Z2 → Y
- 交絡変数 (confounding variables, confounders): XとYの両者の原因となる変数
  - Z1



交絡変数とバックドア経路 (2)

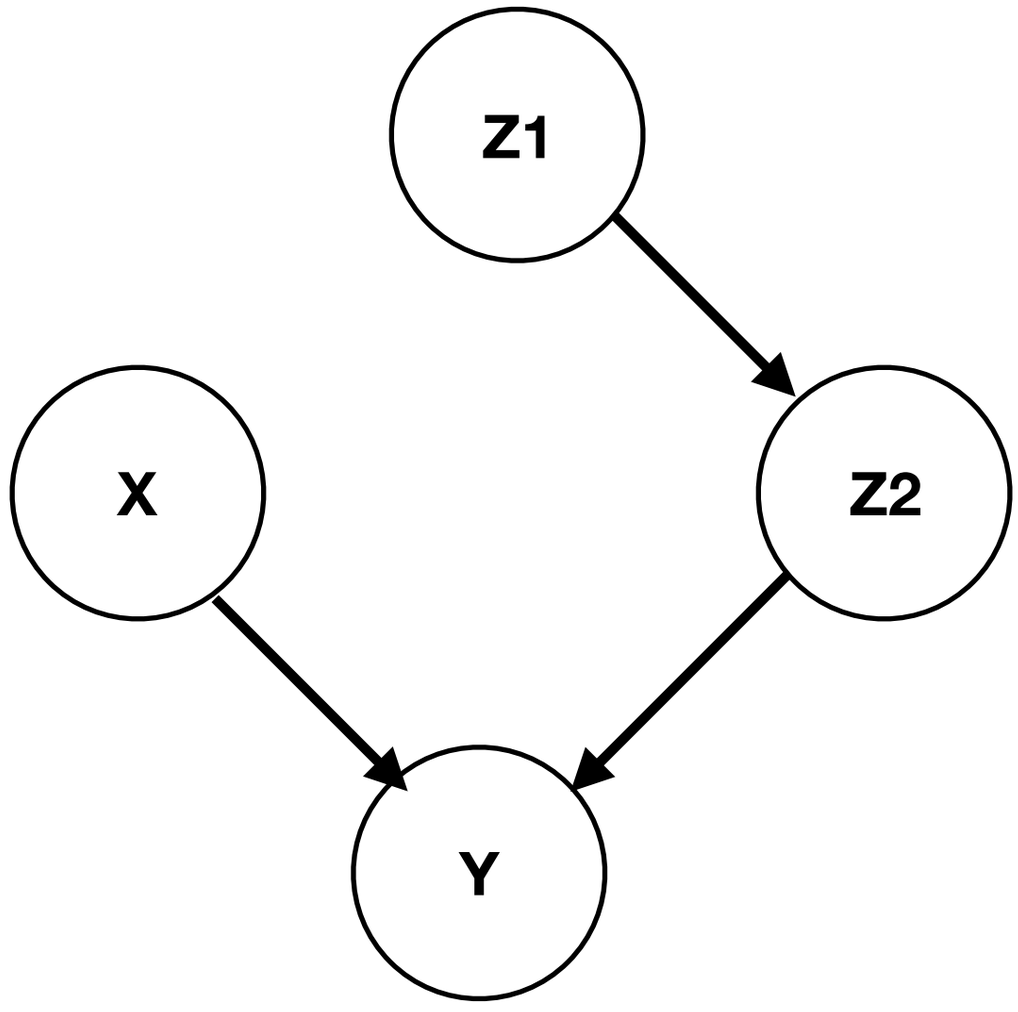
- 右の図にバックドア経路は存在しない
  - X ←Z1 → Z2 はバックドア経路ではない！



- 交絡変数はない
  - Z1 は交絡ではない

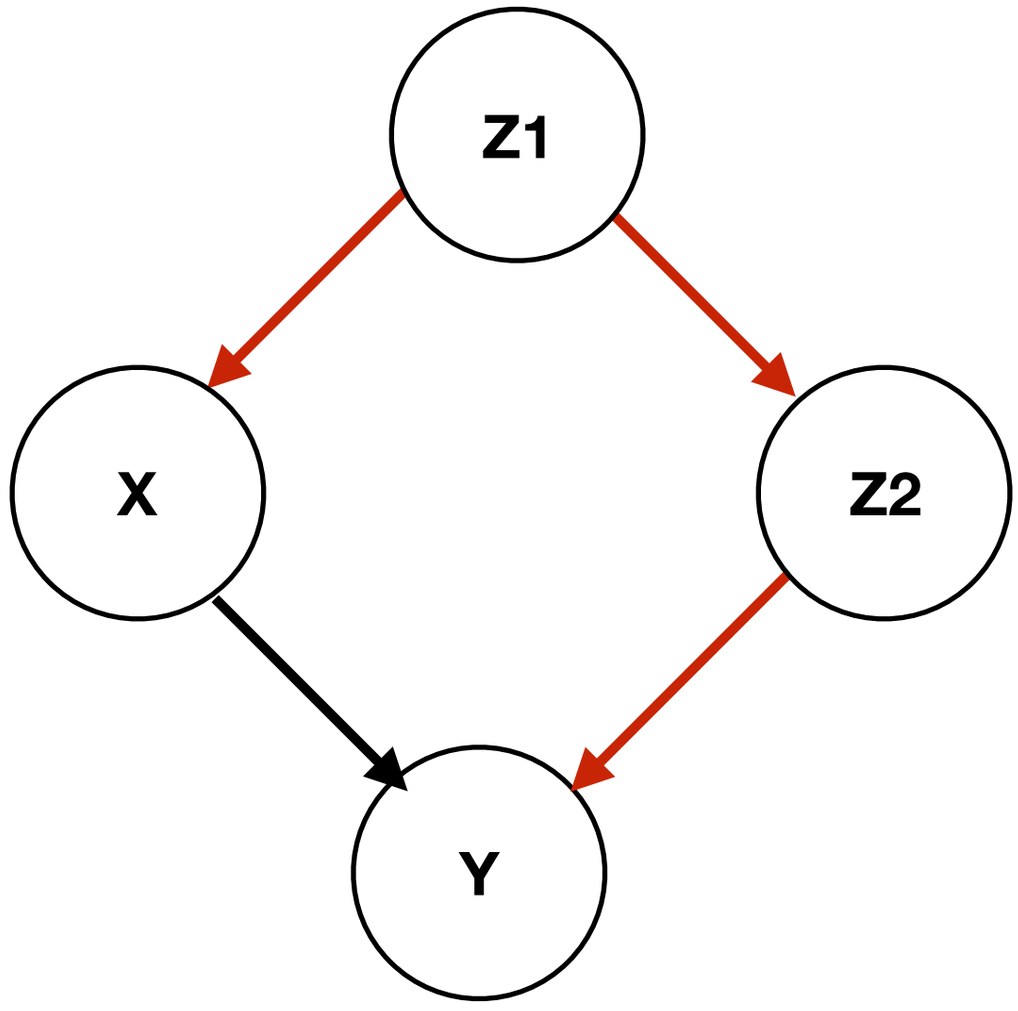
交絡変数とバックドア経路 (1)

- 右の図にバックドア経路は存在しない
  - Z1 → Z2 → 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：
  - X ←Z1 → Z2 → Y
- バックドアを閉じたい
- どうすればいい？

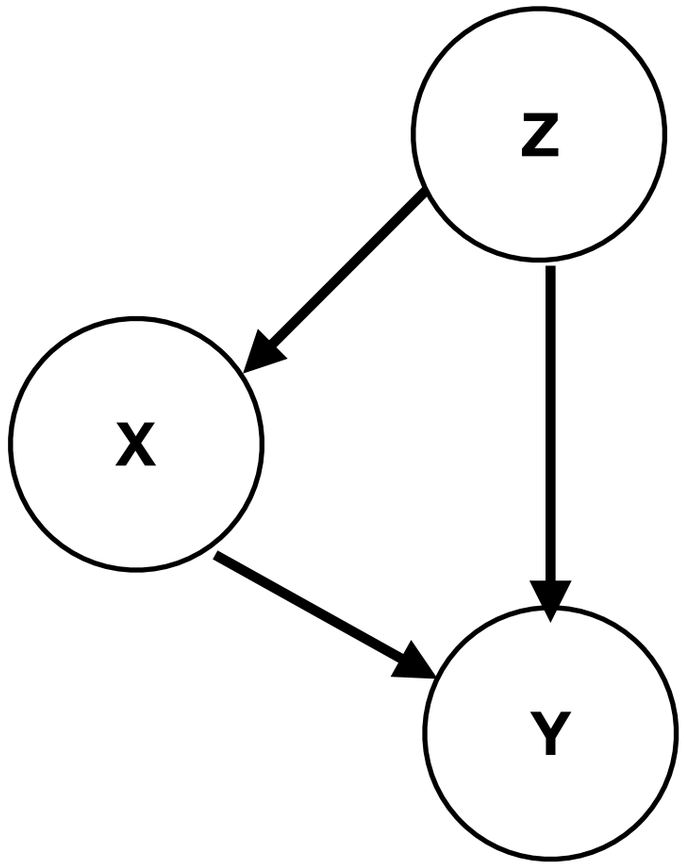


変数 X, Y, Z の関係

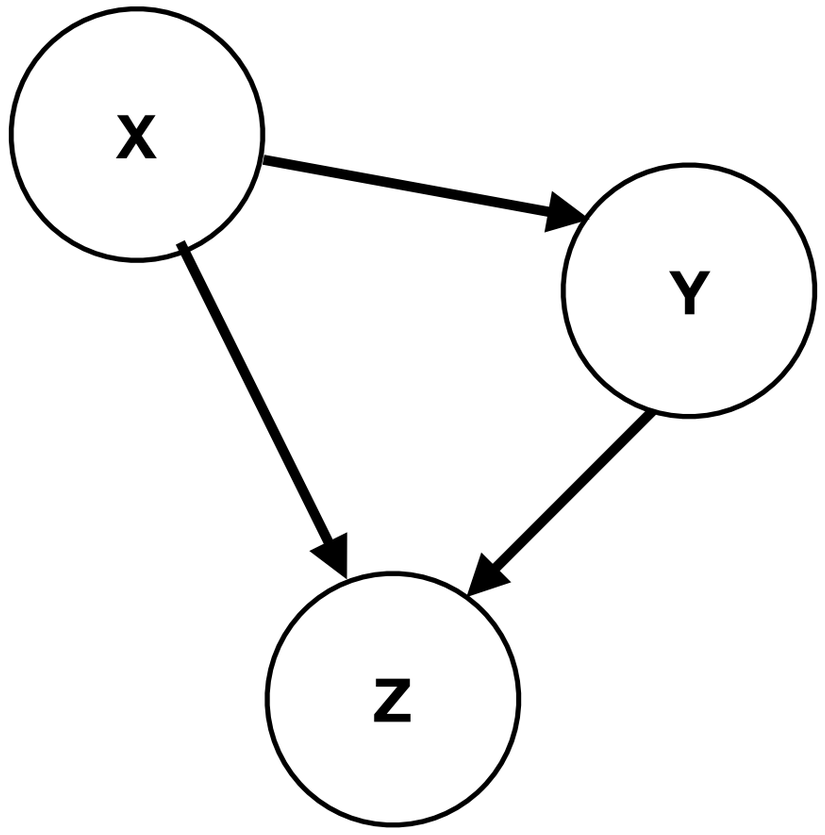
- Yは結果、Xは原因とする
- 3つの可能性

1. Z は X と Y の交絡変数 (confounder) である
2. Z は X と Y の合流点 (collider) である
3. Z は X と Y の媒介変数（mediator, 中間因子）である

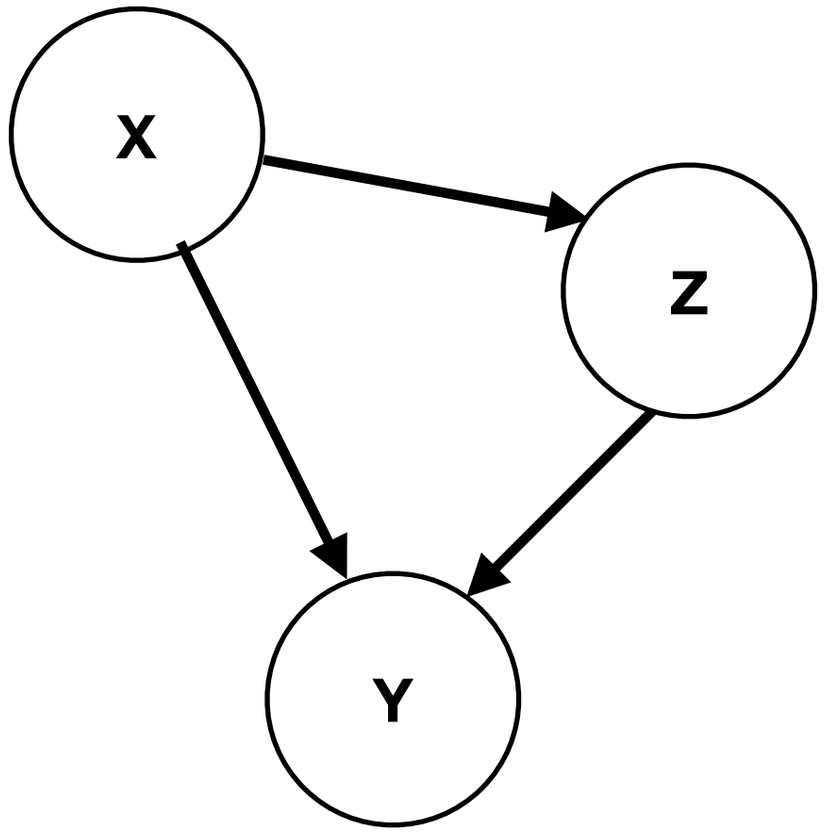
交絡変数 Z



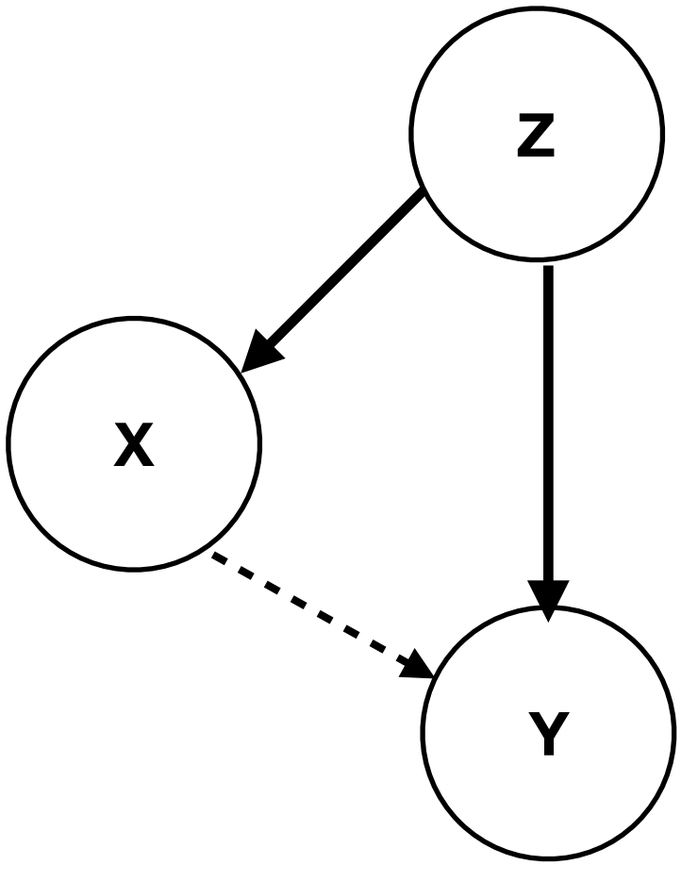
合流点 Z

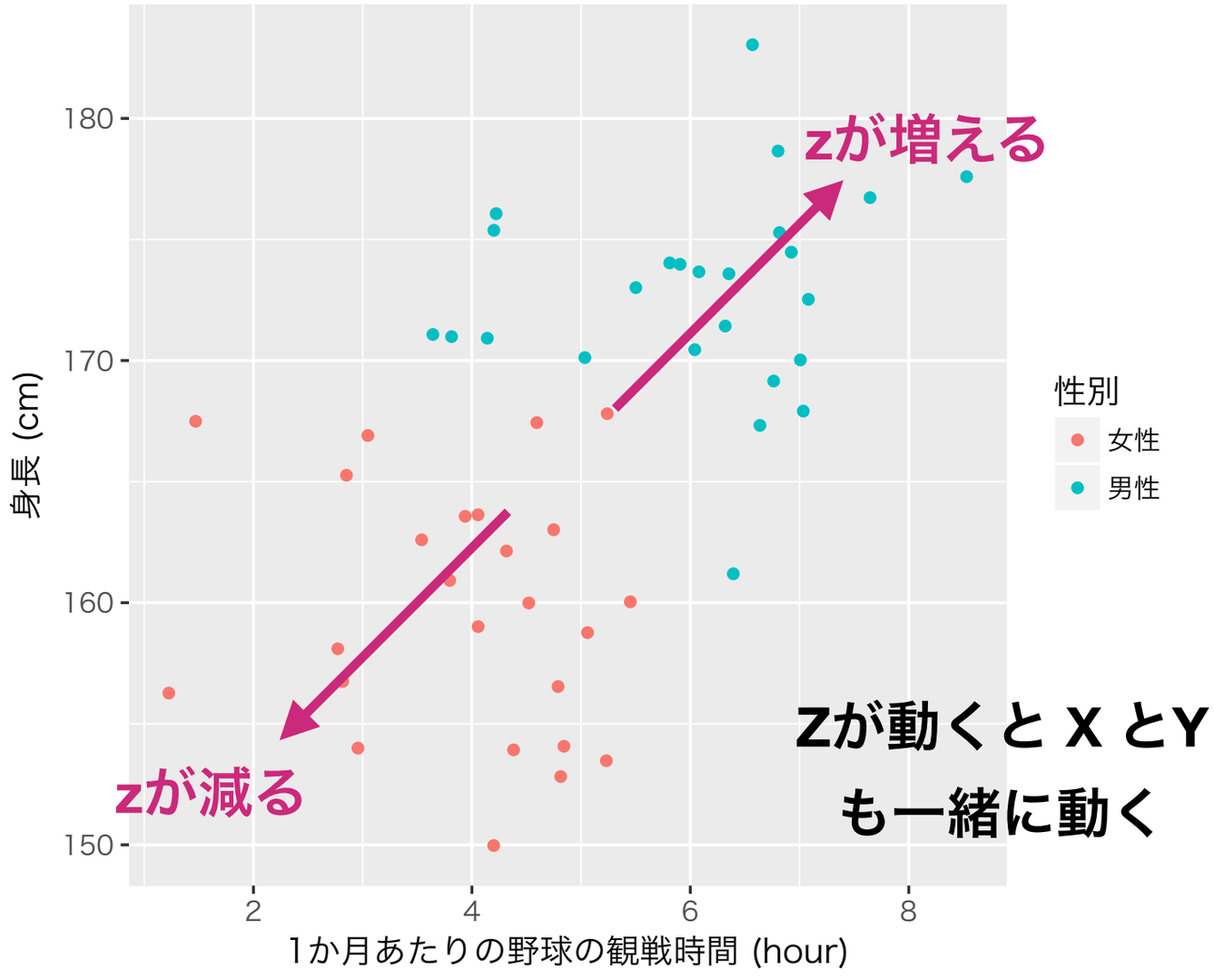


媒介変数 Z



Zが交絡変数のとき

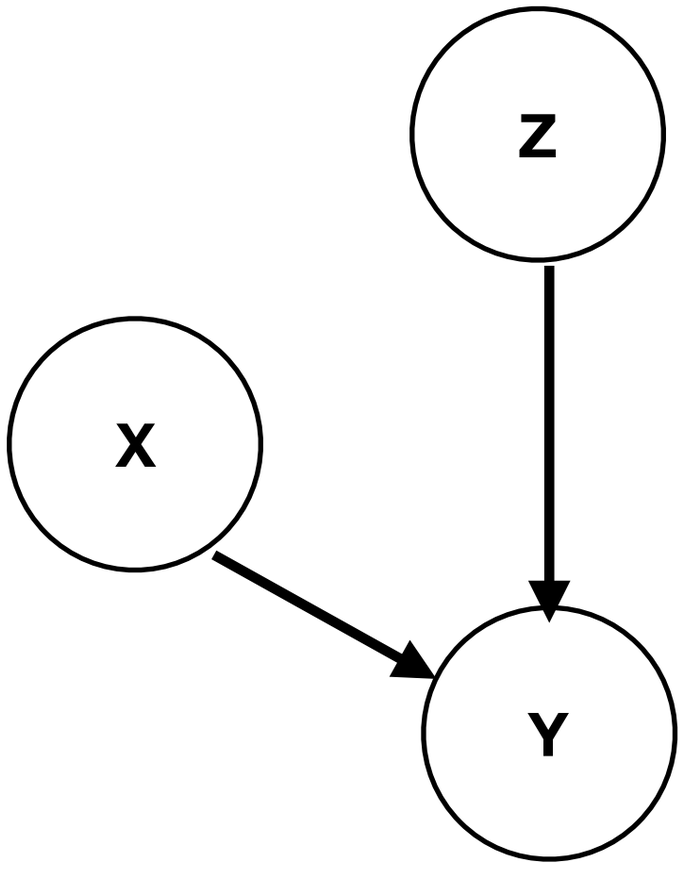


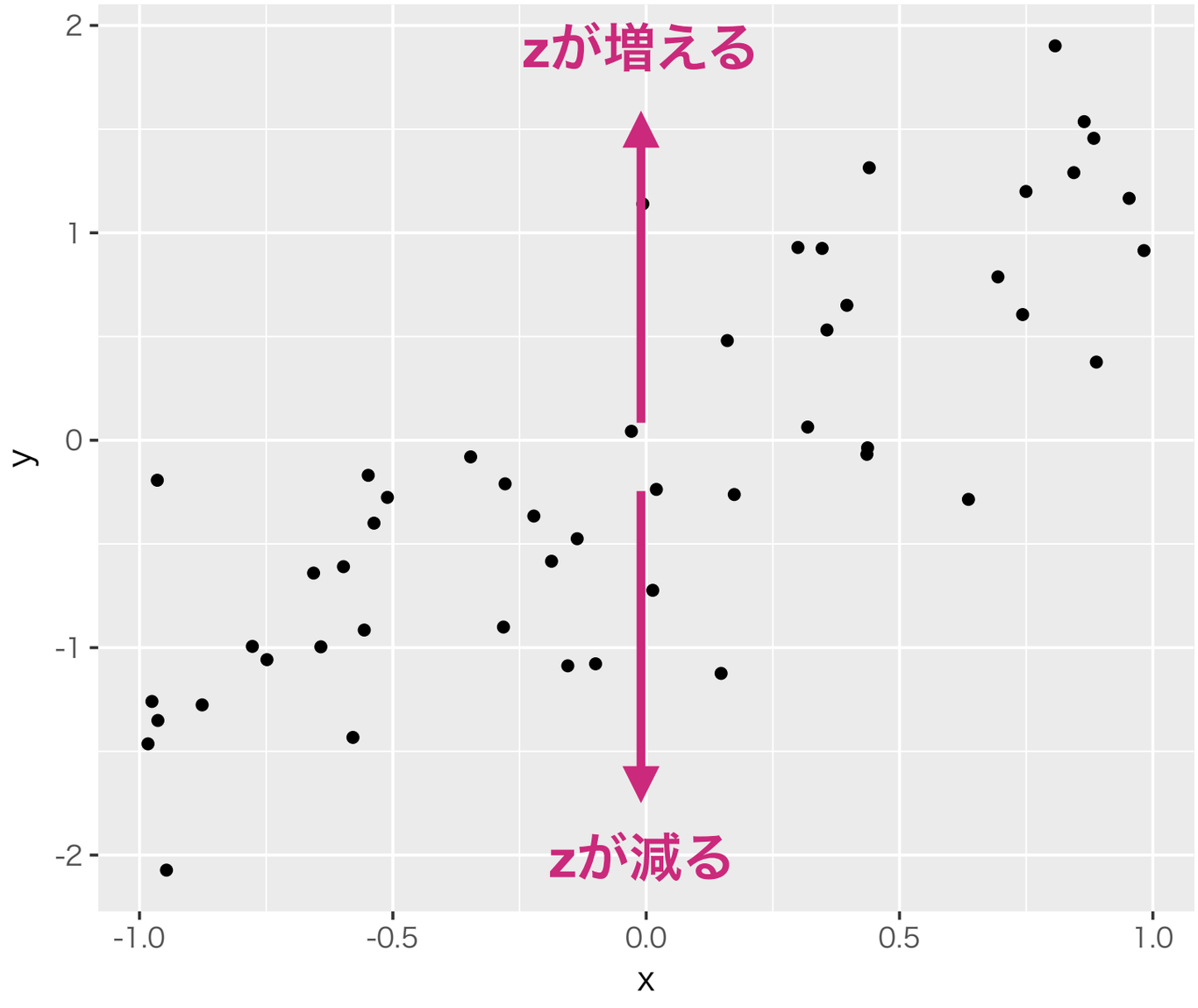


- バックドアが開いていると、Zの変化がXとYの変化を同時に引き起こす

- Y を Xだけに回帰すると、バイアスが生じる

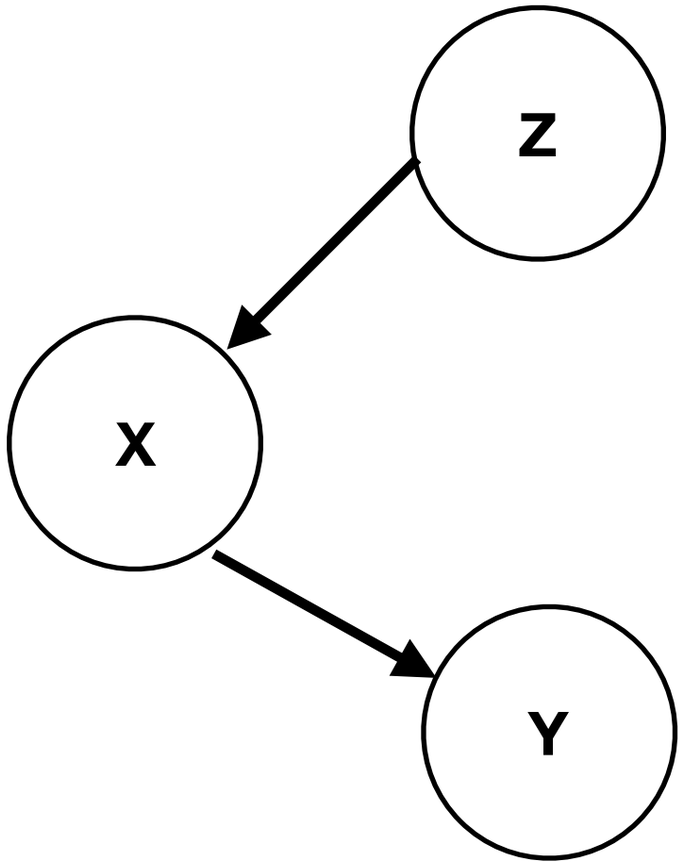
Z が交絡ではない場合 (1)

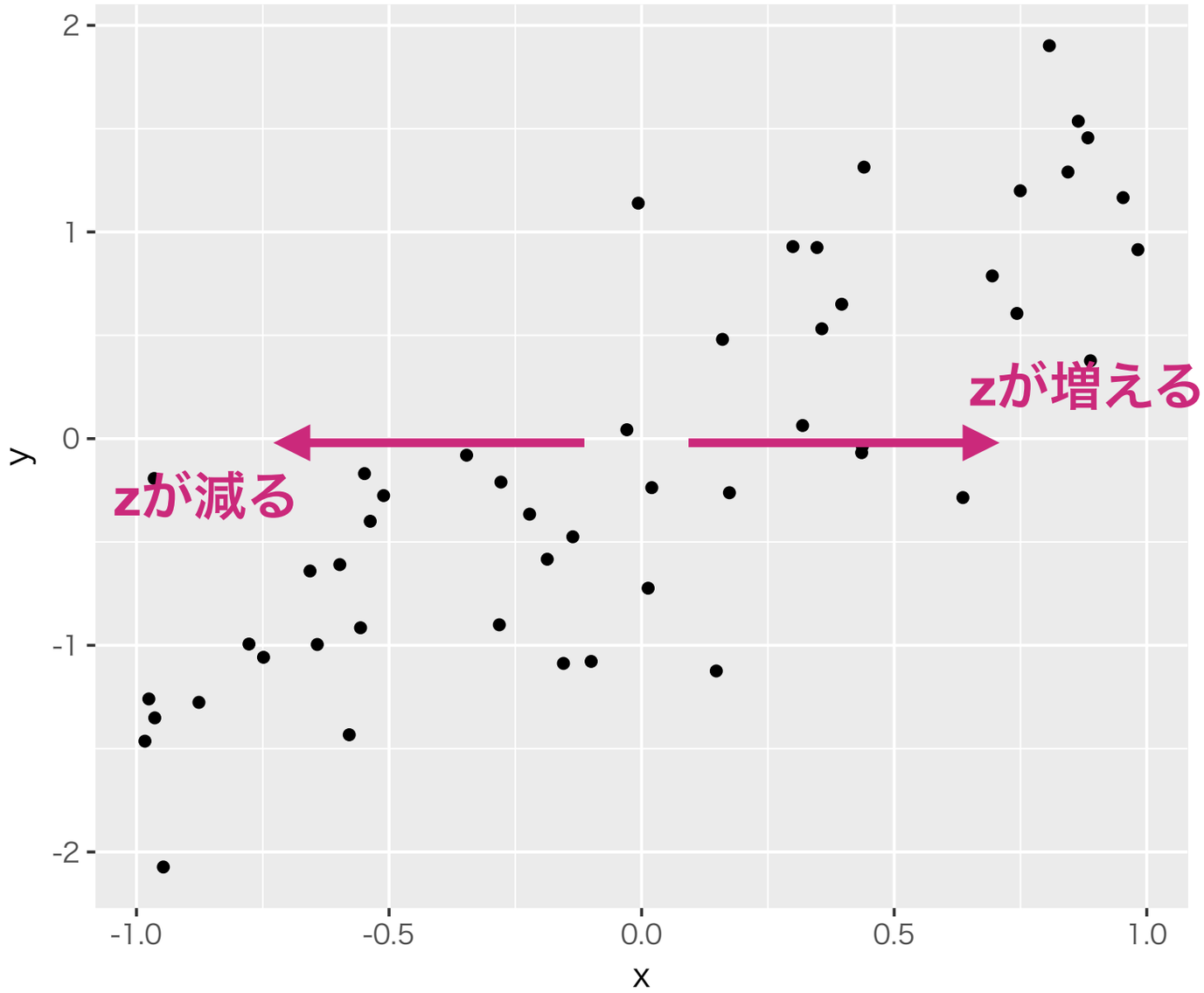




- Zの変化は、Xの変化には影響しない

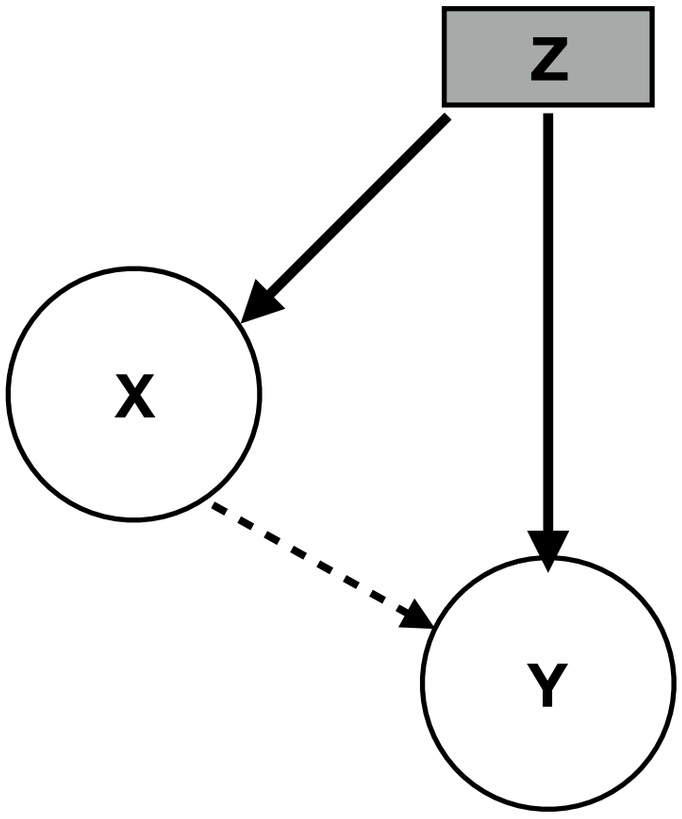
Z が交絡ではない場合 (2)

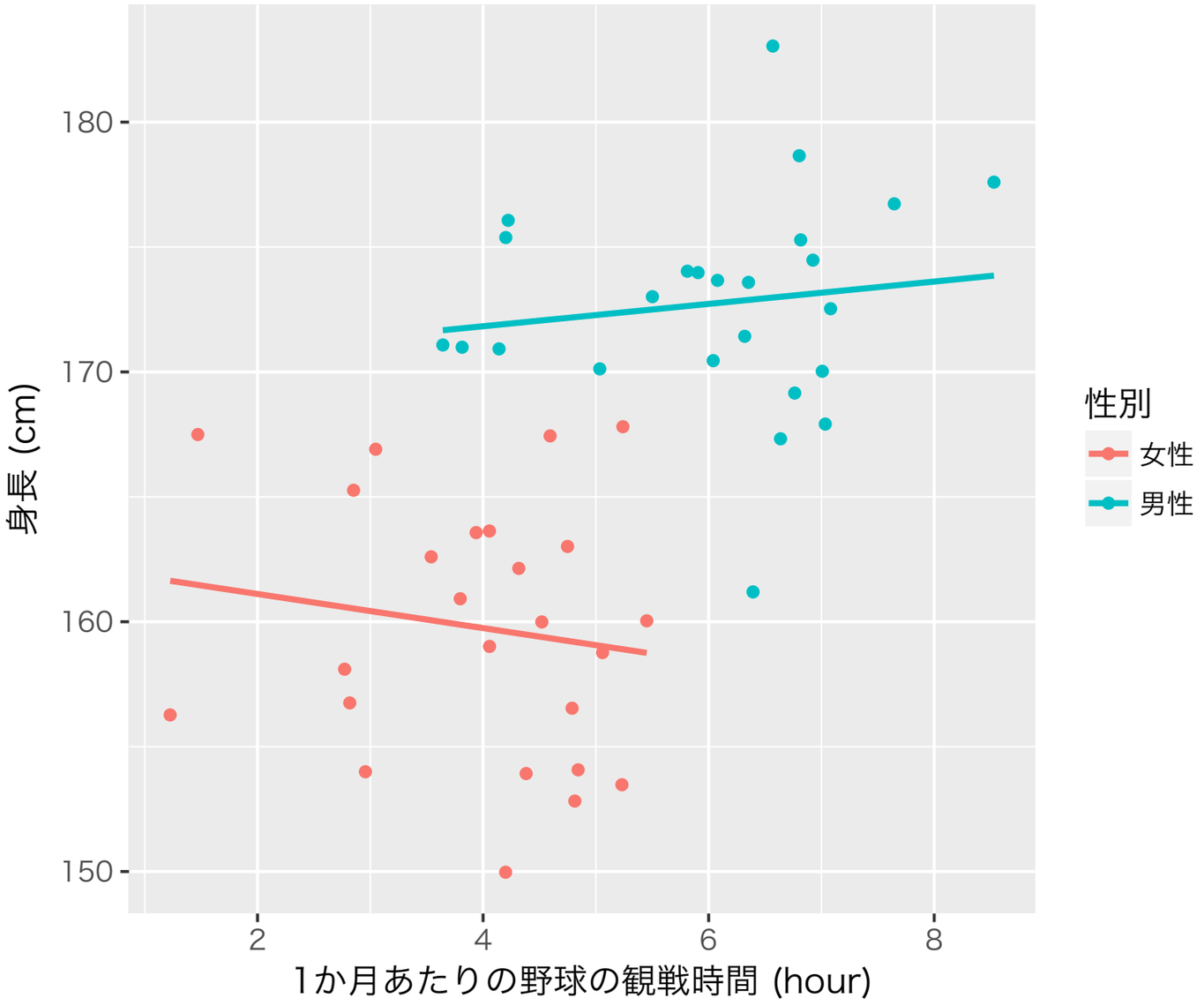




- Zの変化は、Yの変化には影響しない

バイアスを取り除くには？

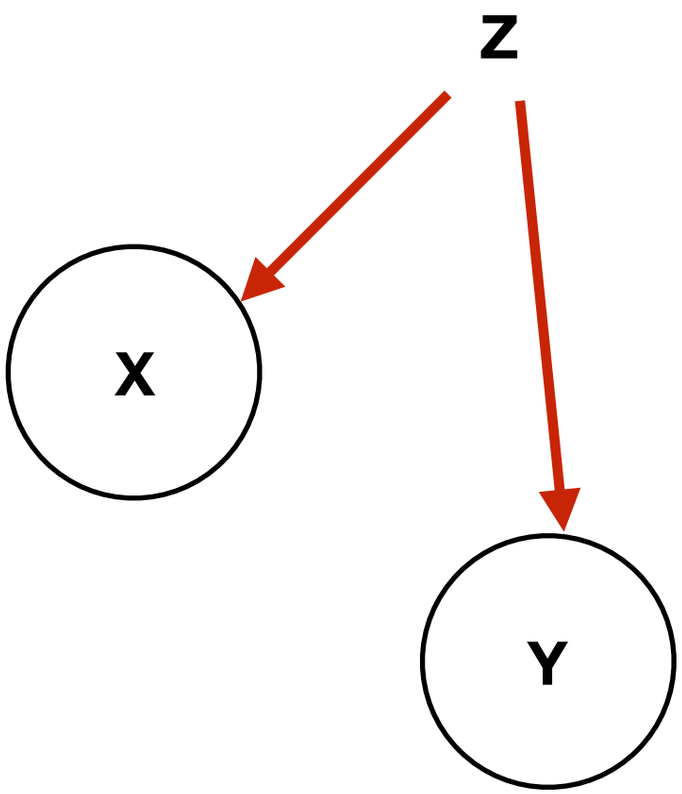




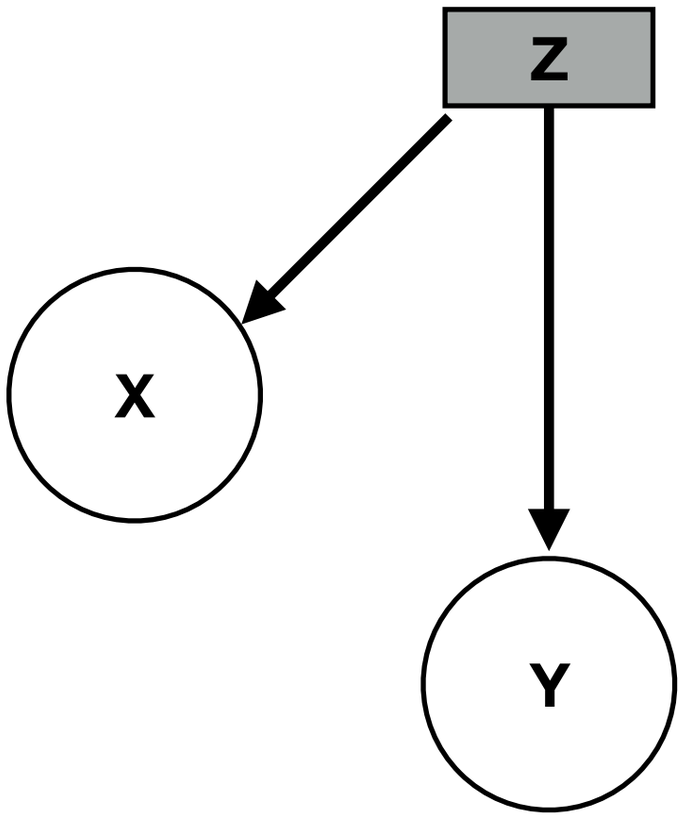
- Zの値を「固定」すれば
  - Zを統制した重回帰分析

バックドアを閉じる

Zなしの回帰



Zを含む回帰



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

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

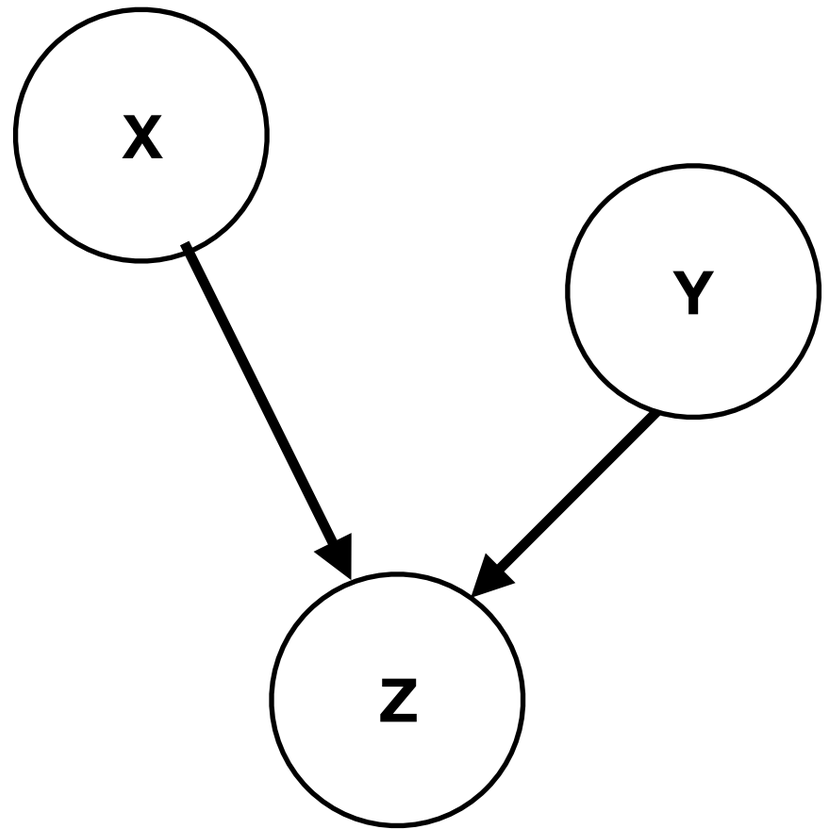
回帰分析における交絡変数の扱い方

- 交絡は統制（コントロール）せよ！

- 交絡を統制すれば、バイアスを防げる

- 交絡を統制し損ねると、欠落変数バイアス (omitted variable bias) が生じる

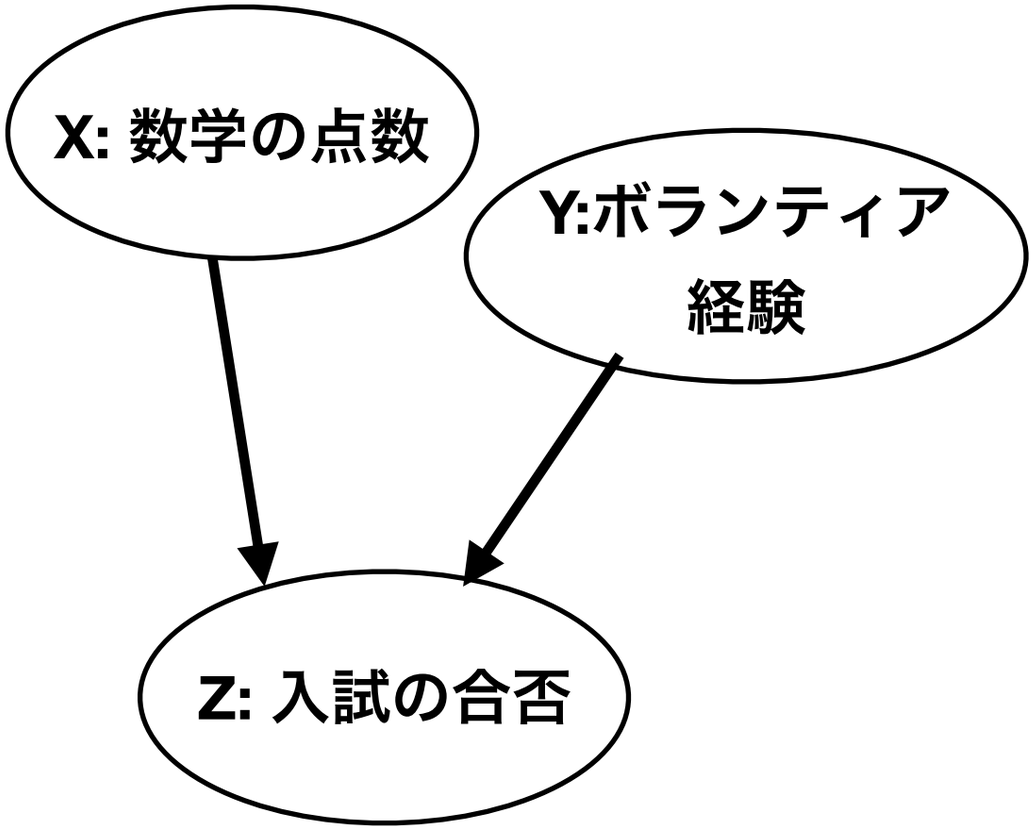
Zが合流点のとき

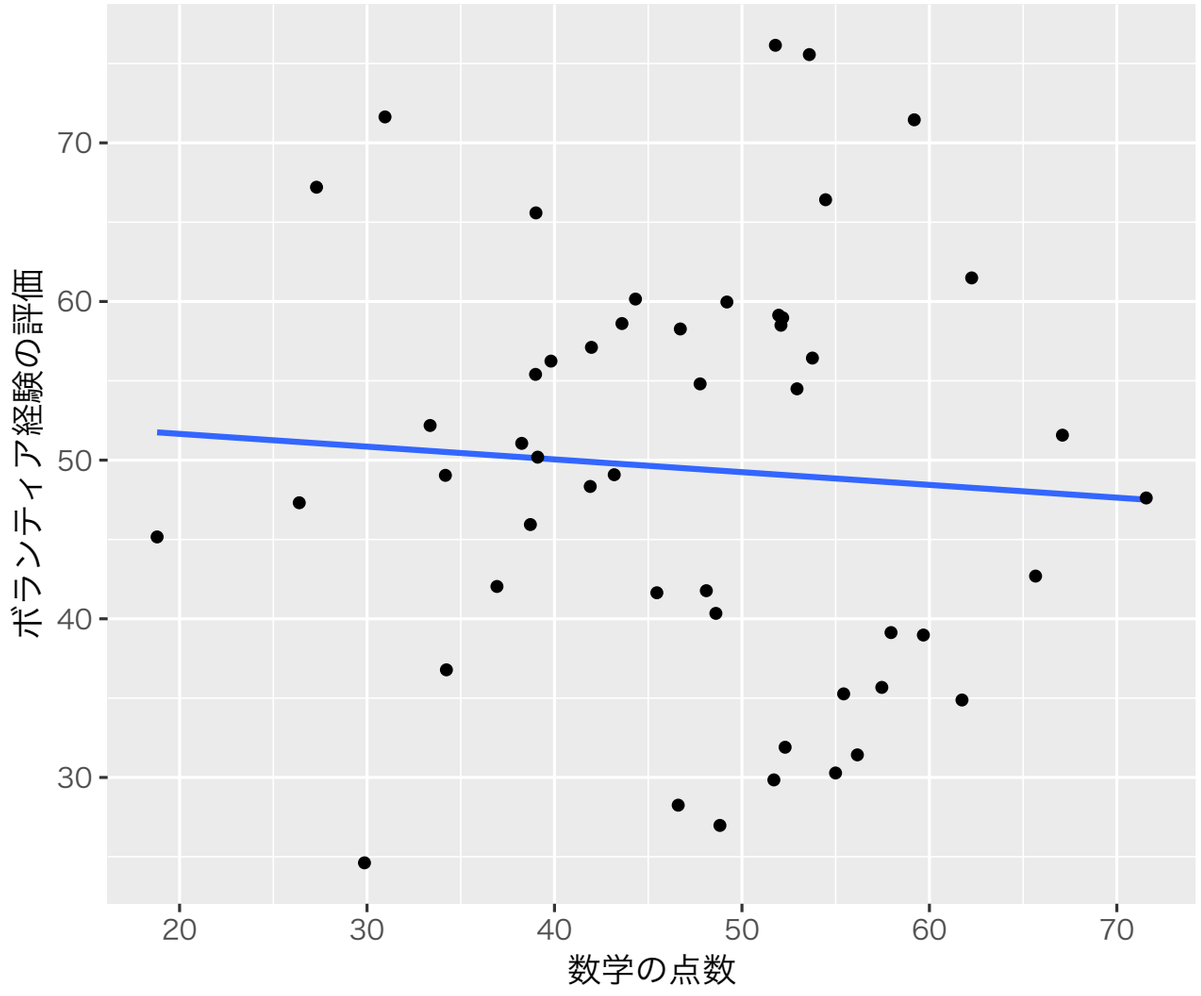


- Zを無視した単回帰で、XのYに対する因果効果を推定できる

合流点を統制すると何が起きる？ (1)

例：アメリカ合衆国の大学入試

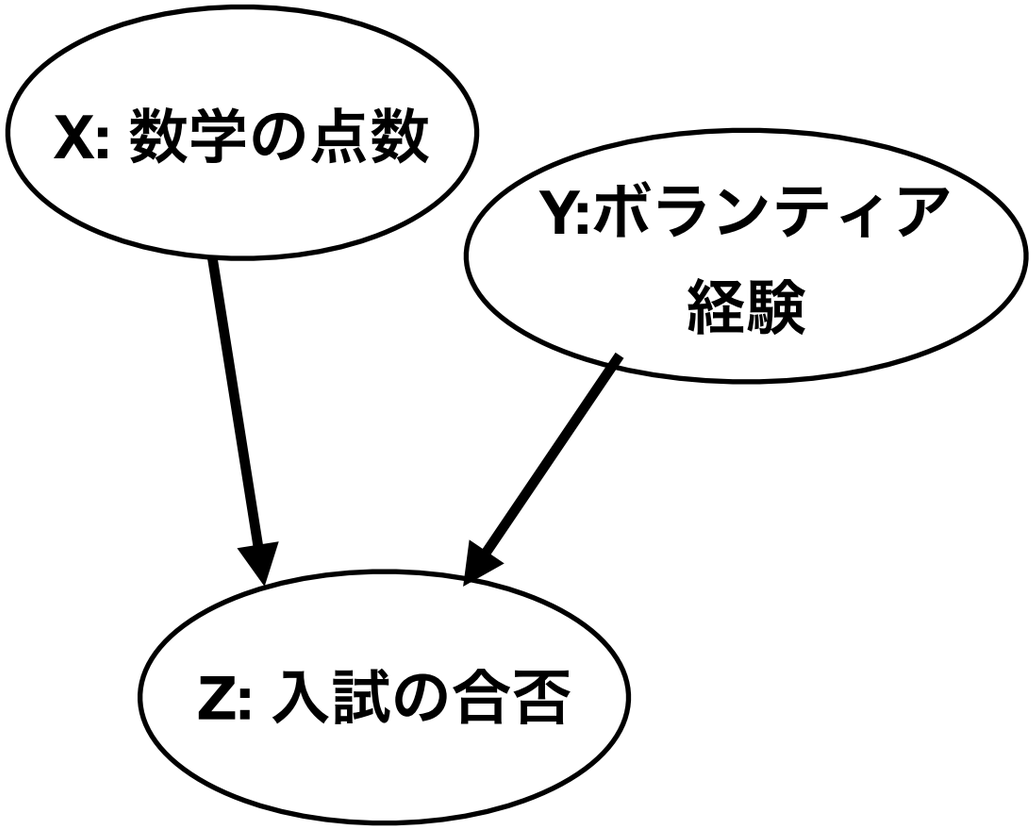


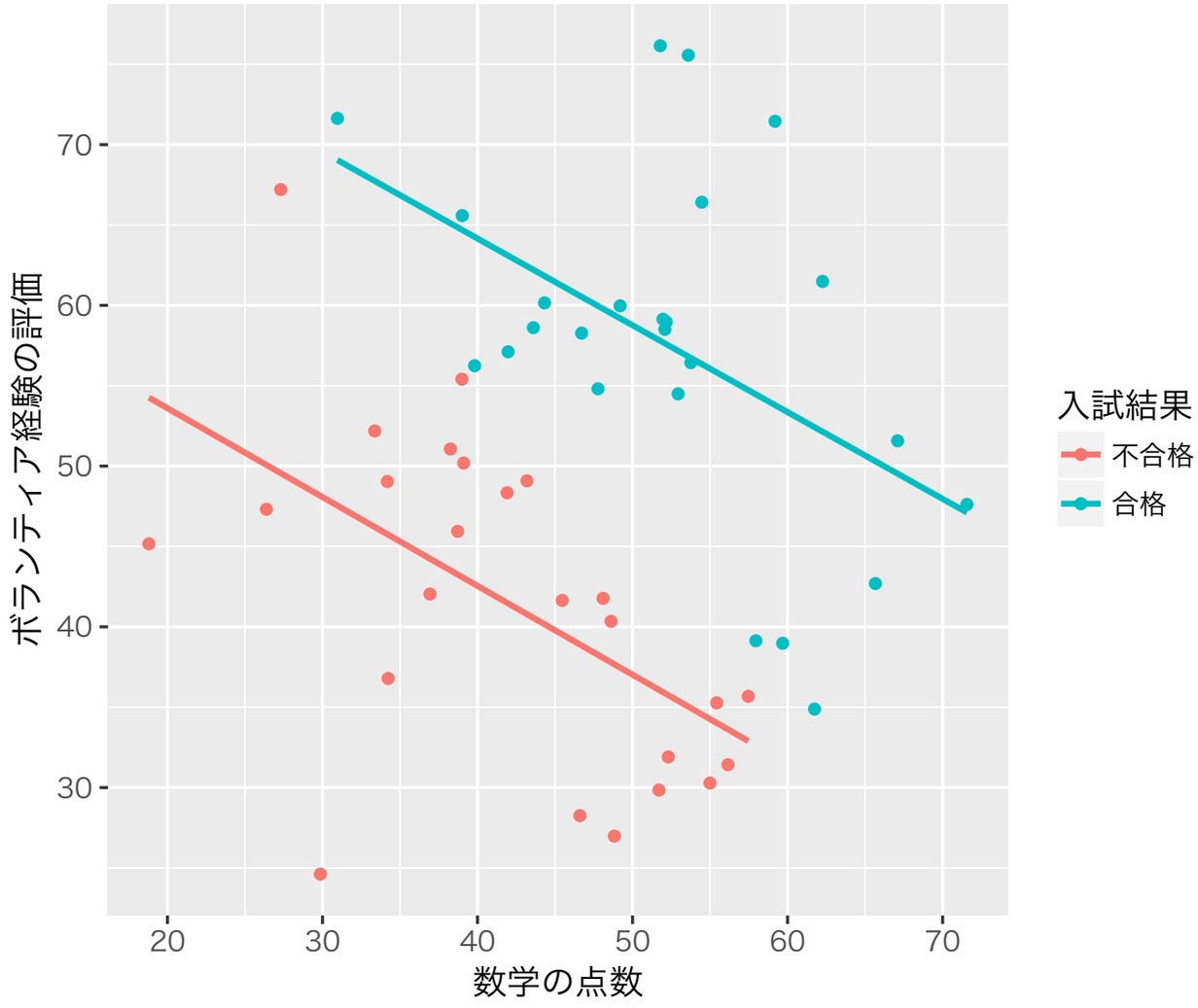


- 入試の合否は、数学の点数とボランティア経験の評価によって決まる（架空のデータ）

合流点を統制すると何が起きる？ (2)

例：アメリカ合衆国の大学入試

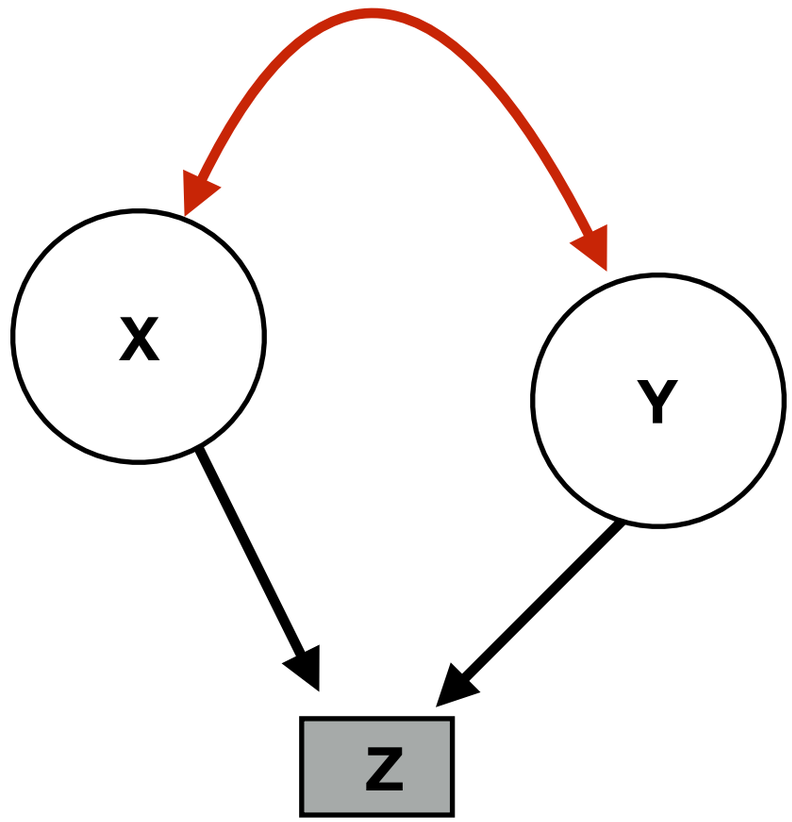




- 合流点Zを統制すると、重回帰で因果効果ではない効果を捉えてしまう

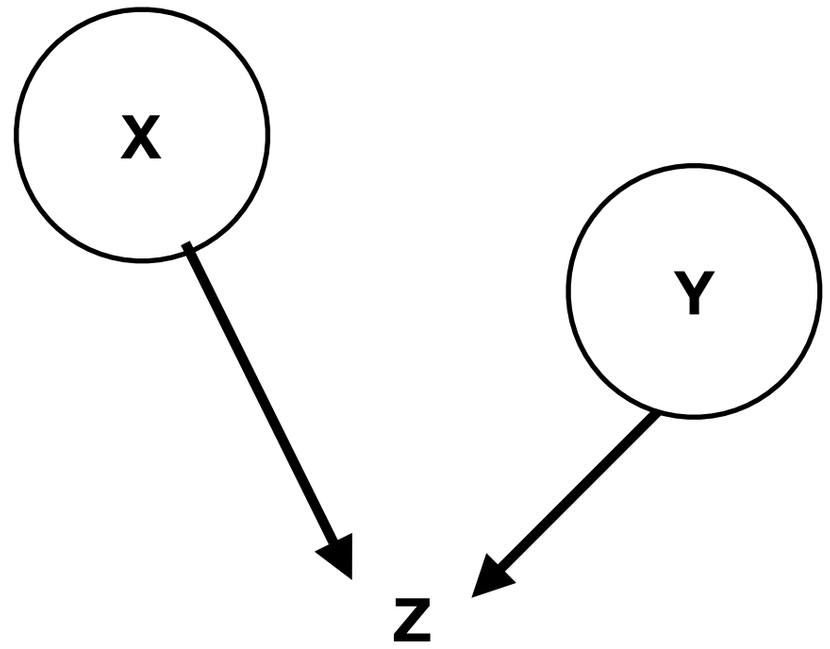
合流点とバックドア経路

Zを含む回帰



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

Zを含まない回帰

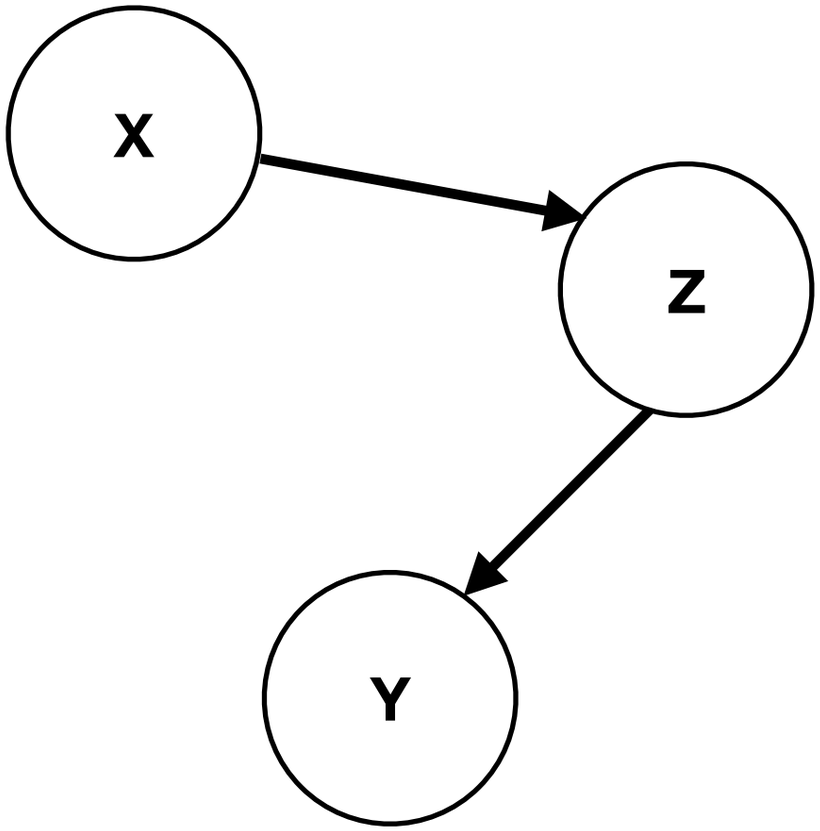


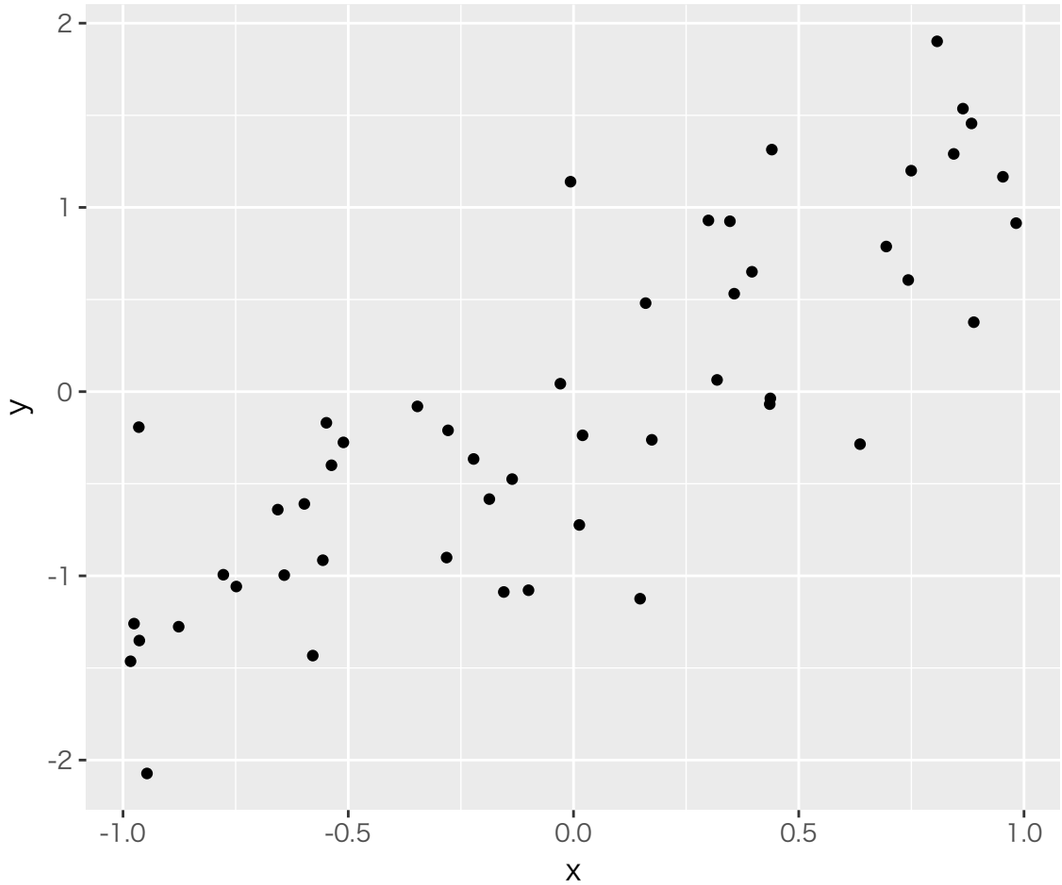
バックドアは存在しない

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

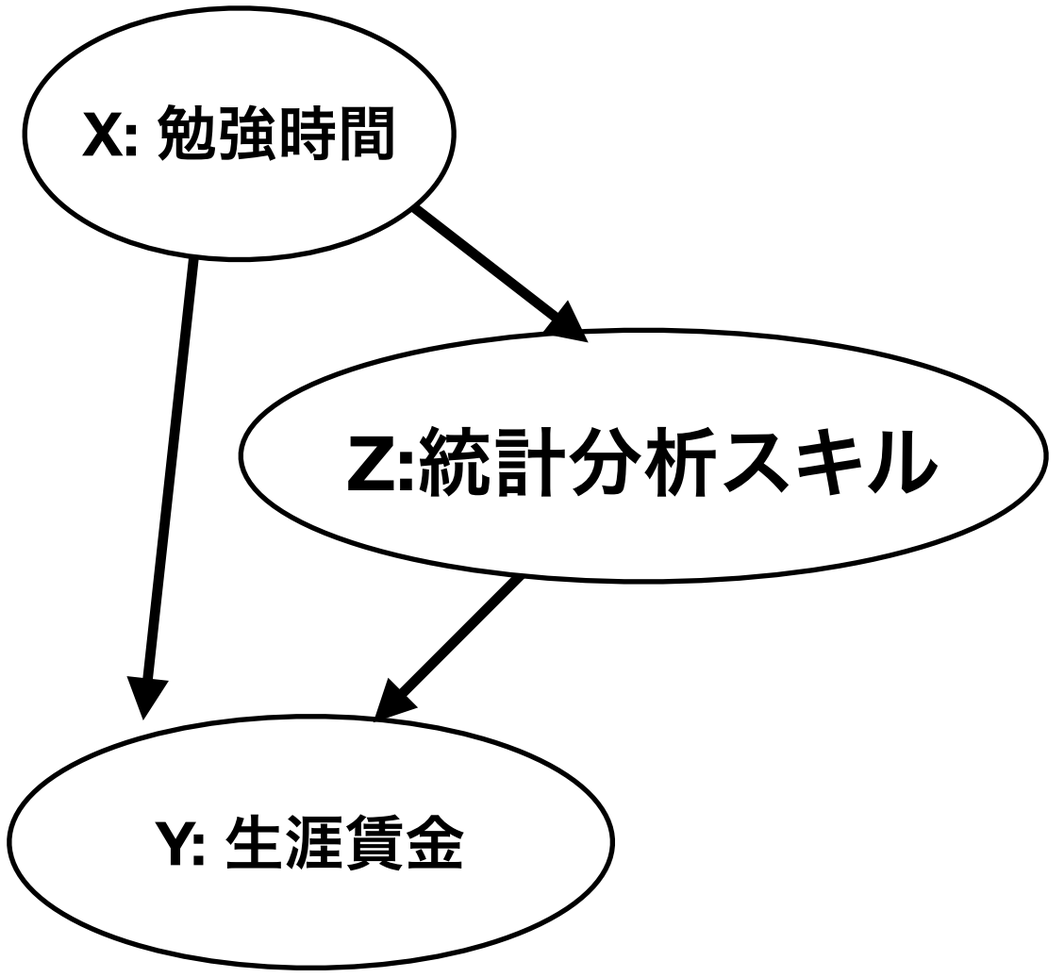
Zが媒介変数のとき

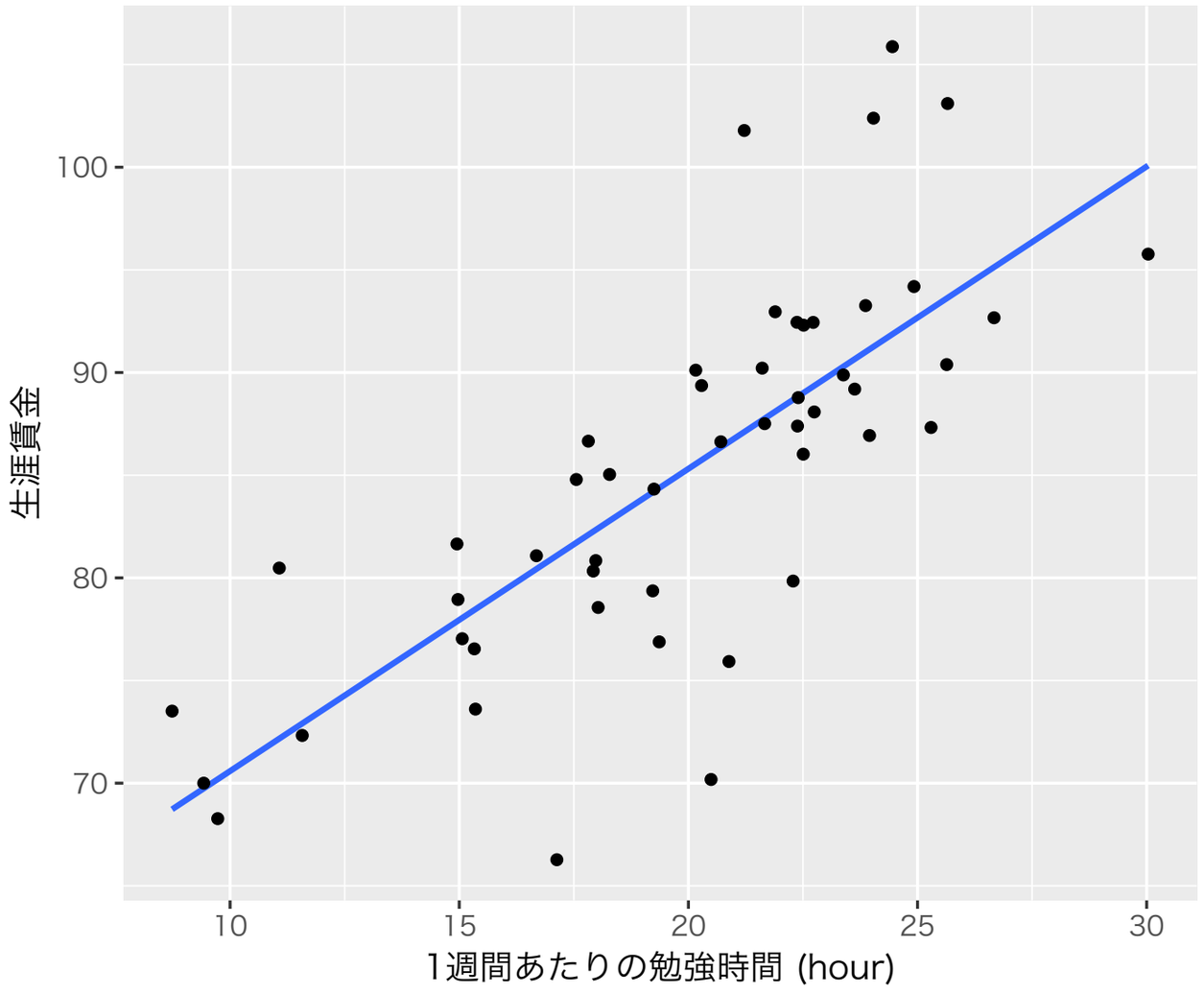




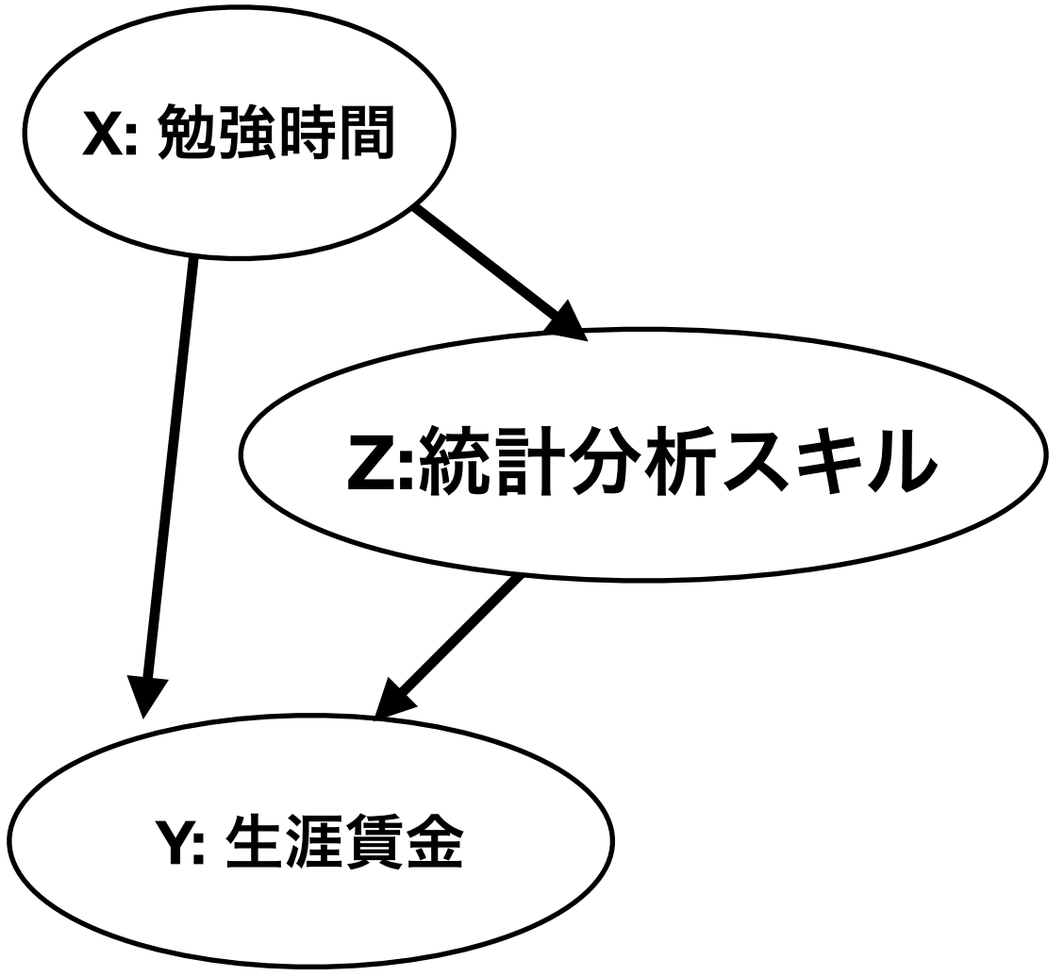
- Zを含まない単回帰モデルで、因果効果を推定できる

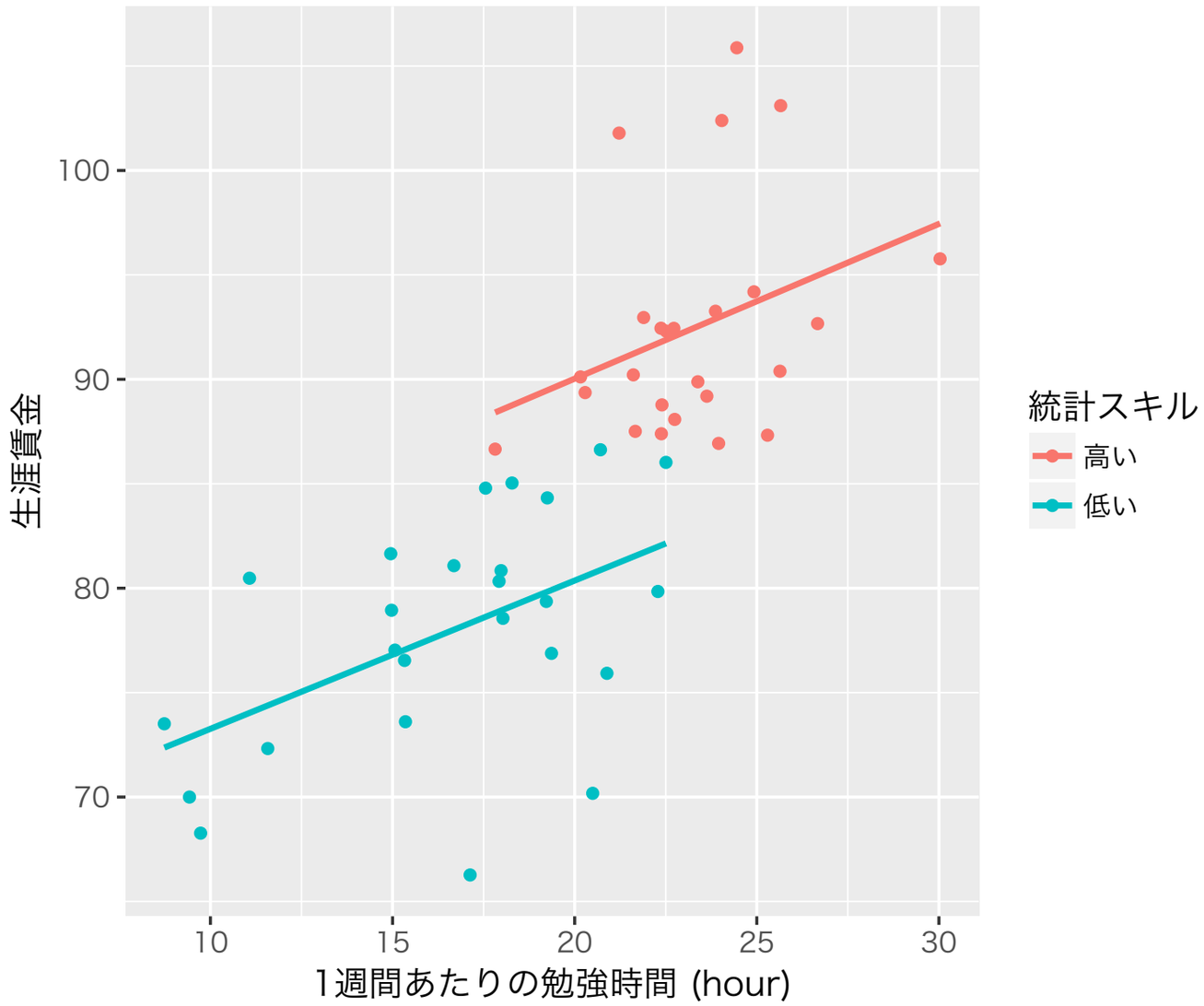
媒介変数を統制すると何が起きる？ (1)





媒介変数を統制すると何が起きる？ (2)





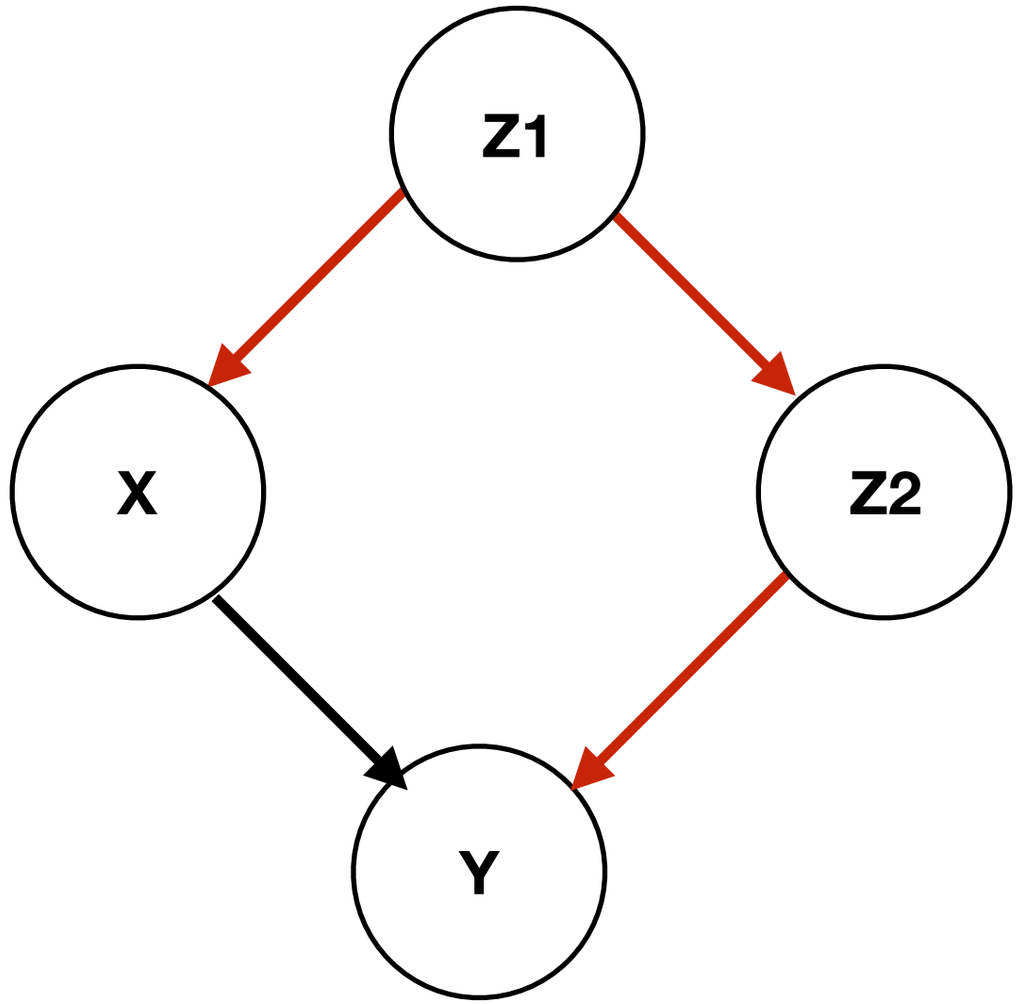
- 媒介変数Zを統制すると、XからYの経路が塞がれてしまう
  - 因果効果が過小評価される

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

- バックドアを閉じればよい
  - Z1を統制する
  - Z2を統制する
  - Z1 と Z2 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きる？

- 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