Rで学ぶ
傾向スコア解析入門

@yokkuns: 里　洋平

2011.09.24 第17回R勉強会@東京（#TokyoR）

AGENDA

- 自己紹介
- 傾向スコア解析
  - 実験出来るデータ
  - 実験出来ないデータ
  - 共変量調整
  - 傾向スコア推定
  - 傾向スコアを用いた調整
- Rによる実行
- 最後に

AGENDA

- 自己紹介
- 傾向スコア解析
  - 実験出来るデータ
  - 実験出来ないデータ
  - 共変量調整
  - 傾向スコア推定
  - 傾向スコアを用いた調整
- Rによる実行
- 最後に

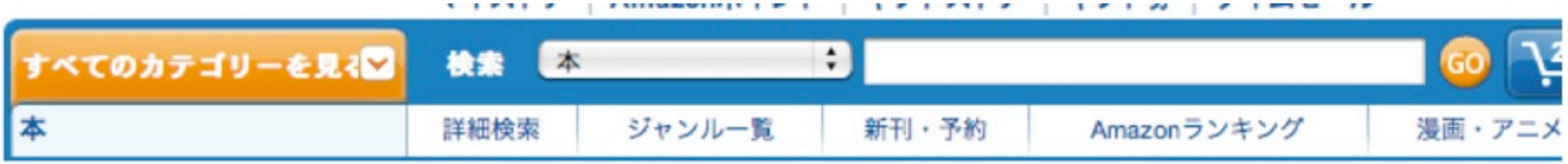
自己紹介

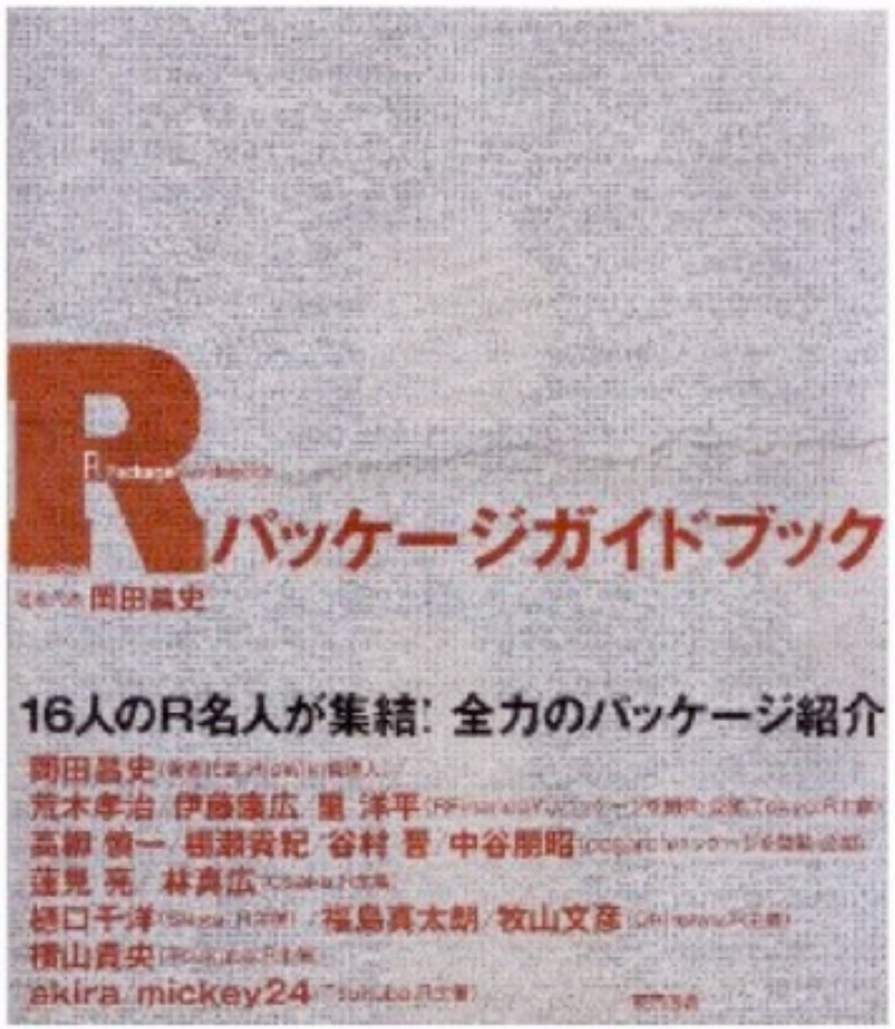
- ID : yokkuns
- 名前 : 里 洋平
- データマイニングエンジニア
- 統計解析 パターン認識
  機械学習 データマイニング
  NLP 金融工学 などを勉強中

Tokyo.Rの主催者

ご参加ありがとうございます！

Rパッケージ本執筆





Rパッケージガイドブック [単行本]

岡田 昌史 (著), 荒木 孝治 (その他), 伊藤 康広 (その他), 里 洋平 (その他), 高柳 慎一 (その他), 棚瀬 貴紀 (その他), 谷村 晋 (その他), 中谷 朋昭 (その他), 蓮見 亮 (その他), 林 真広 (その他), 樋口 千洋 (その他), 福島 真太朗 (その他), 牧山 文彦 (その他), 横山 貴央 (その他), akira (その他), mickey24 (その他)

この商品の最初のレビューを書き込んでください。 いいね (6)

価格： ￥3,990 通常配送無料 詳細

通常2~4週間以内に発送します。 在庫状況について

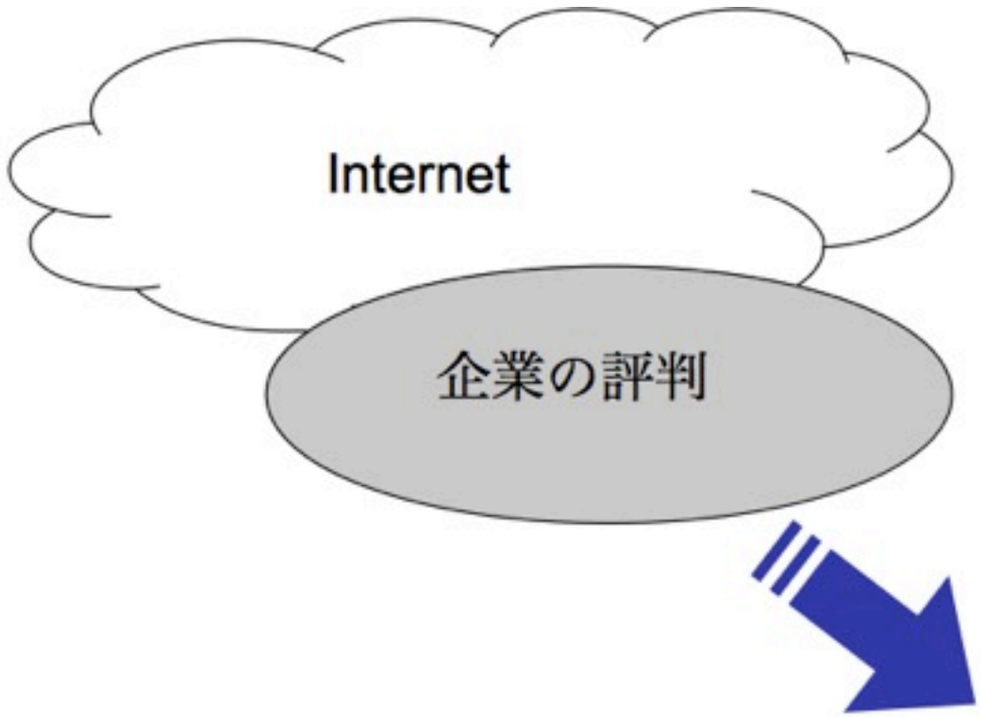
この商品は、Amazon.co.jp が販売、発送します。 ギフトラッピングを利用できます。

中古品1点¥ 6,133より

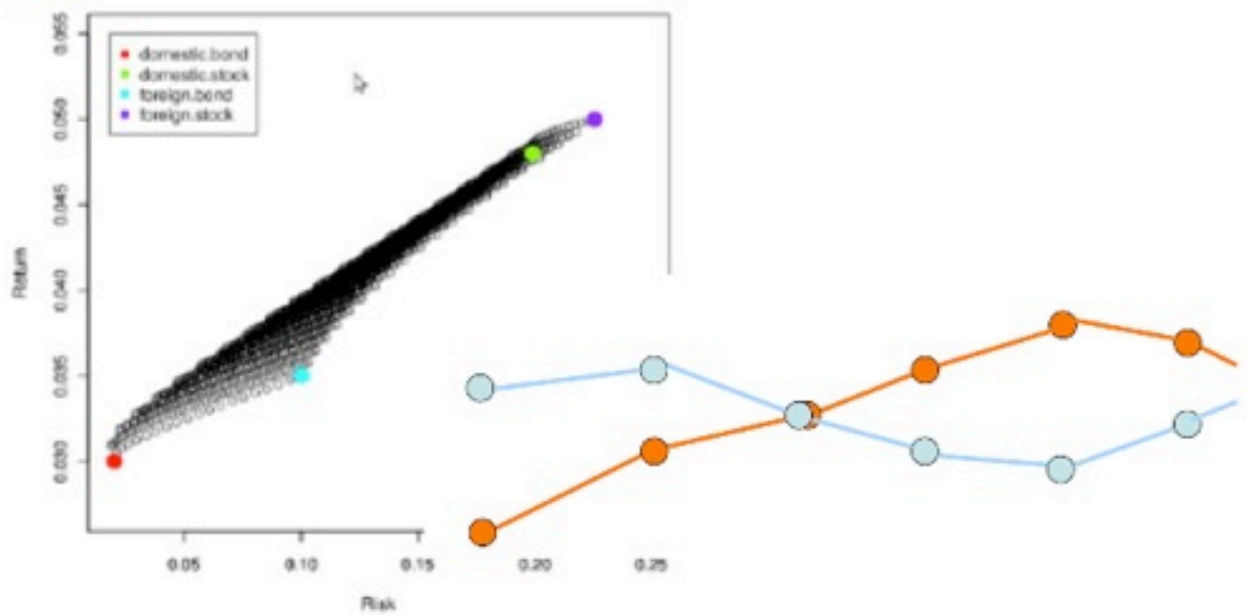
動画レコメンド


SF家電
洗濯機篇
関連オススメ動画リスト
次のリストを見る
絞り込み動画
前へ
5 / 30
かしこい
家電の選び方
SF家電
電子レンジ2編
この出演者は
だれ？
男の家事
（洗濯編）
全件クリア
次へ
視聴履歴を見る
閲覧されてる動画に関連する
おすすめ動画を表示

テキストマイニング+金融工学



Web上にある評判情報から
市場予測！



AGENDA

- 自己紹介
- **傾向スコア解析**
  - **実験出来るデータ**
  - **実験出来ないデータ**
  - **共変量調整**
  - **傾向スコア推定**
  - **傾向スコアを用いた調整**
- Rによる実行
- 最後に

傾向スコア解析

実験出来ないデータの因果関係を解析する

3歳神話：子供は3歳までは母親の元で育つ方が社会性・知能発達が向上する

1～3歳まで母親の側

8歳時点の
社会性得点・知能検査
を比較

1～3歳に保育園

傾向スコア解析

実験出来ないデータの因果関係を解析する

3歳神話：子供は3歳までは母親の元で育つ方が社会性･知能発達が向上する

1～3歳まで母親の側

◇親の学歴

8歳時点の
社会性得点･知能検査
を比較

1～3歳に保育園

◇親の学歴

傾向スコア解析

実験出来ないデータの因果関係を解析する

3歳神話：子供は3歳までは母親の元で育つ方が社会性･知能発達が向上する

1～3歳まで母親の側

◇親の学歴

◇親の収入

8歳時点の
社会性得点･知能検査
を比較

1～3歳に保育園

◇親の学歴

◇親の収入

傾向スコア解析

実験出来ないデータの因果関係を解析する

3歳神話：子供は3歳までは母親の元で育つ方が社会性･知能発達が向上する

1～3歳まで母親の側

◇親の学歴
◇親の収入
◇親の職業

8歳時点の
社会性得点･知能検査
を比較

1～3歳に保育園

◇親の学歴
◇親の収入
◇親の職業

傾向スコア解析

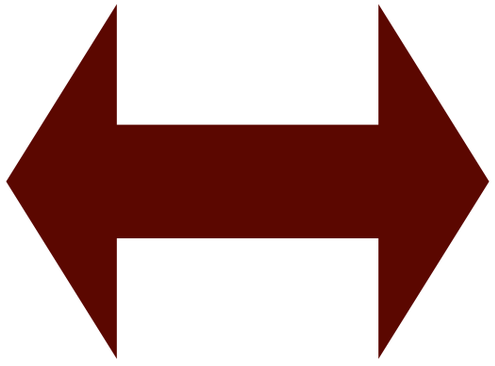
実験出来ないデータの因果関係を解析する

3歳神話：子供は3歳までは母親の元で育つ方が社会性･知能発達が向上する

1～3歳まで母親の側

- ◇親の学歴
- ◇親の収入
- ◇親の職業
- ◇親の教育意欲

8歳時点の
社会性得点･知能検査
を比較



1～3歳に保育園

- ◇親の学歴
- ◇親の収入
- ◇親の職業
- ◇親の教育意欲

傾向スコア解析

実験出来ないデータの因果関係を解析する

3歳神話：子供は3歳までは母親の元で育つ方が社会性･知能発達が向上する

8歳時点の社会性得点･知能検査を比較

1～3歳まで母親の側

- ◇親の学歴
- ◇親の収入
- ◇親の職業
- ◇親の教育意欲

1～3歳に保育園

- ◇親の学歴
- ◇親の収入
- ◇親の職業
- ◇親の教育意欲

親の学歴や収入などが影響している

傾向スコア解析

実験出来ないデータの因果関係を解析する

テレビCMの効果測定

傾向スコア解析

実験出来ないデータの因果関係を解析する

テレビCMの効果測定

傾向スコア解析

実験出来ないデータの因果関係を解析する

テレビCMの効果測定

TVCMを見たグループ

◇購買意欲
◇ライフスタイル

当該商品の
購買量の比較

TVCMを見てないグループ

◇購買意欲
◇ライフスタイル

傾向スコア解析

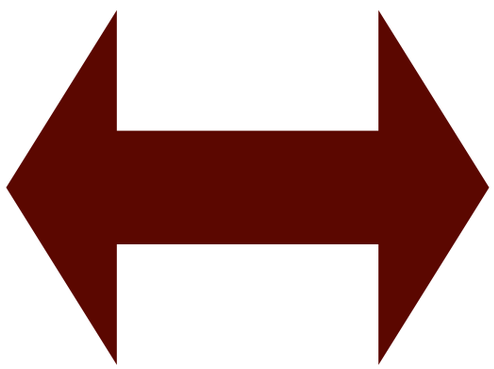
実験出来ないデータの因果関係を解析する

テレビCMの効果測定

TVCMを見たグループ

◇購買意欲
◇ライフスタイル
◇年齢層

当該商品の
購買量の比較



TVCMを見てないグループ

◇購買意欲
◇ライフスタイル
◇年齢層

傾向スコア解析

実験出来ないデータの因果関係を解析する

テレビCMの効果測定

TVCMを見たグループ

◇購買意欲
◇ライフスタイル
◇年齢層
◇所得

当該商品の
購買量の比較

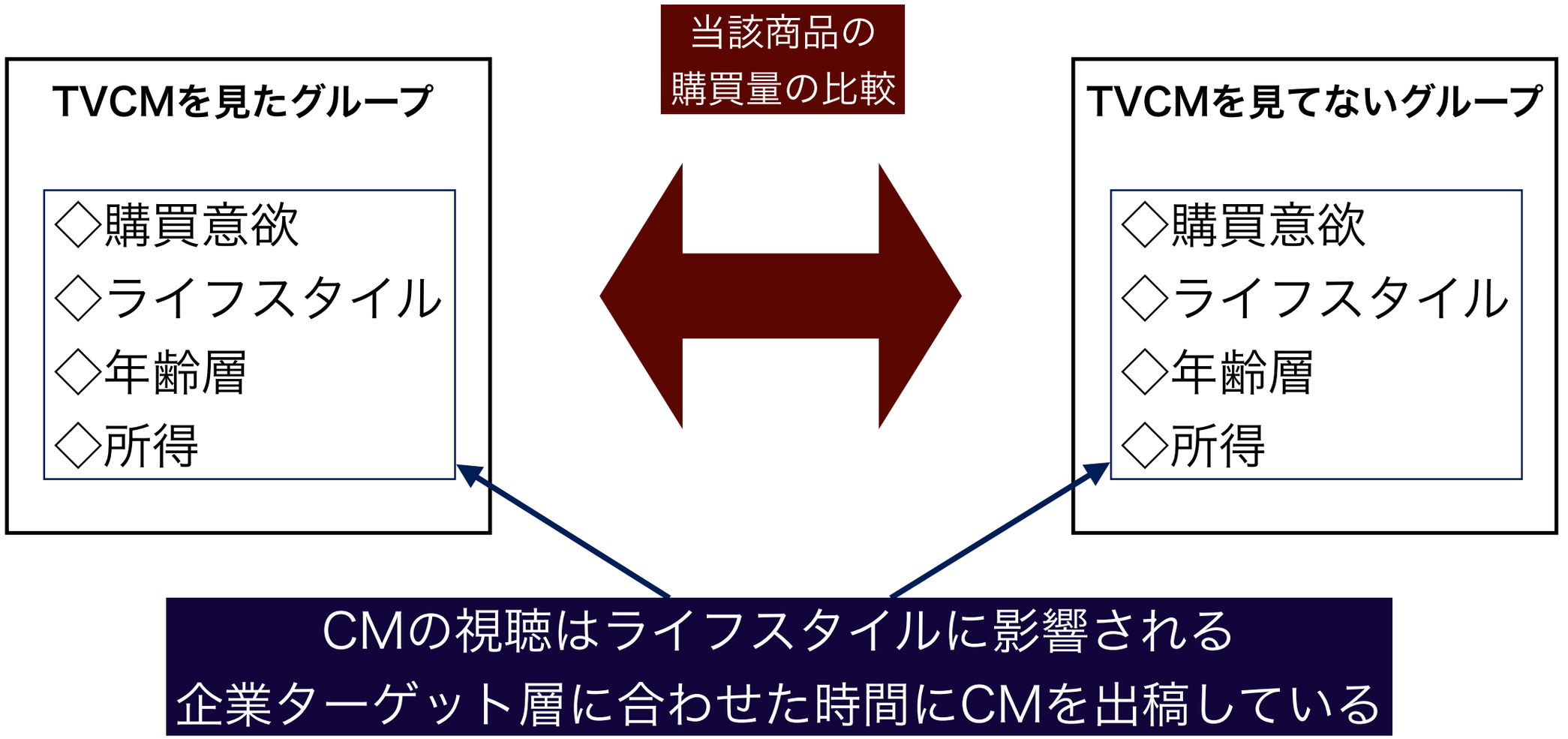
TVCMを見てないグループ

◇購買意欲
◇ライフスタイル
◇年齢層
◇所得

傾向スコア解析

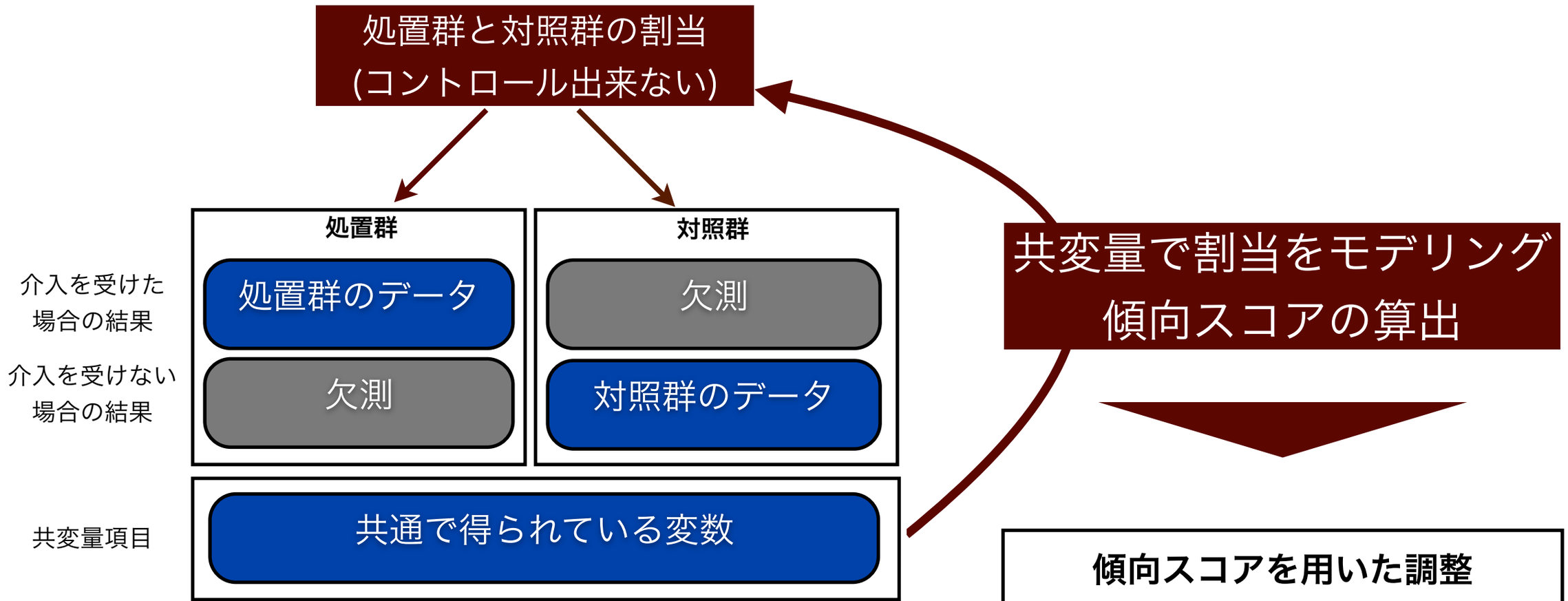
実験出来ないデータの因果関係を解析する

テレビCMの効果測定



傾向スコア解析

実験出来ないデータの因果関係を解析する

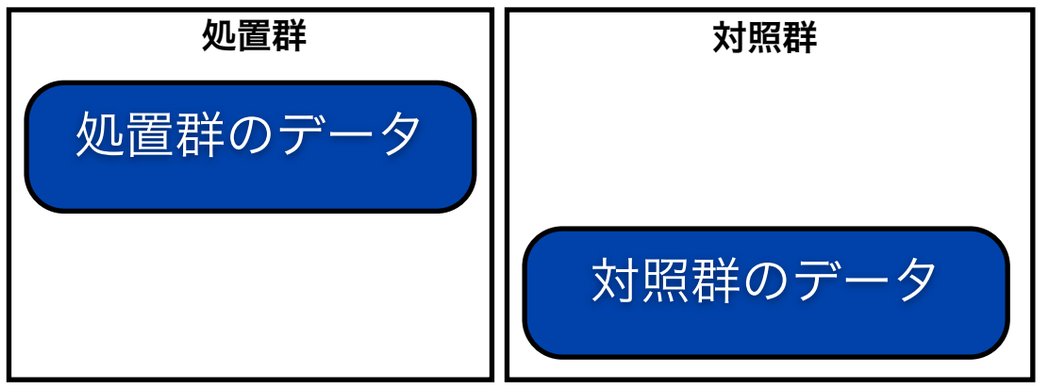


傾向スコアを用いた調整

◇マッチング
◇層別解析
◇共分散分析

実験出来るデータ

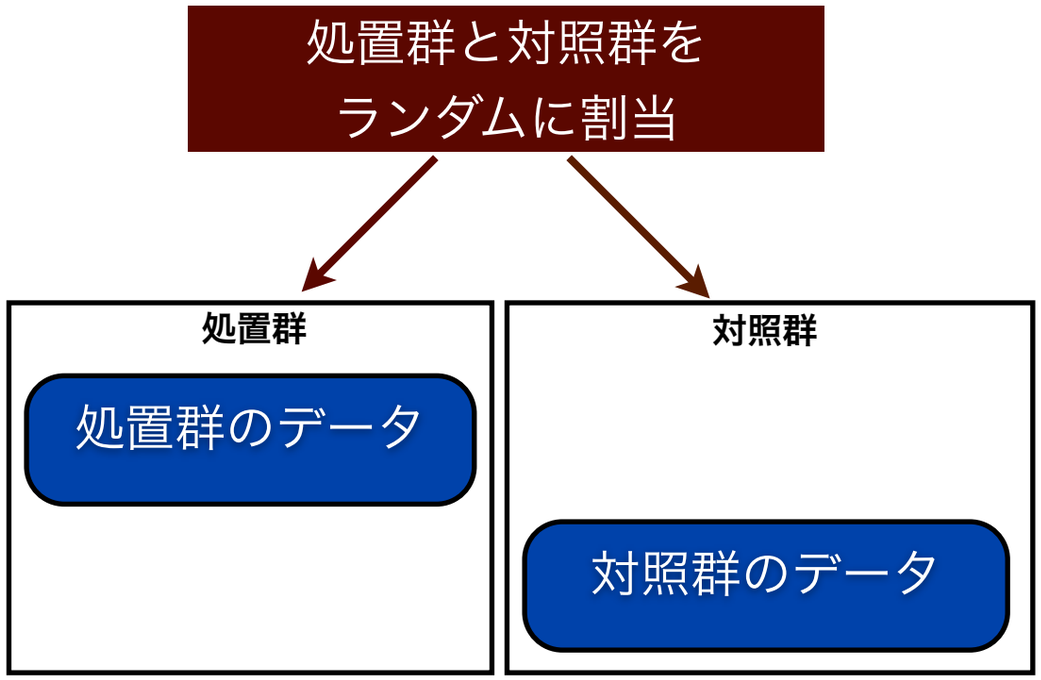
因果効果は単純な処置群と対照群の差になる



因果効果 = 処置群の平均 - 対照群の平均

実験出来るデータ

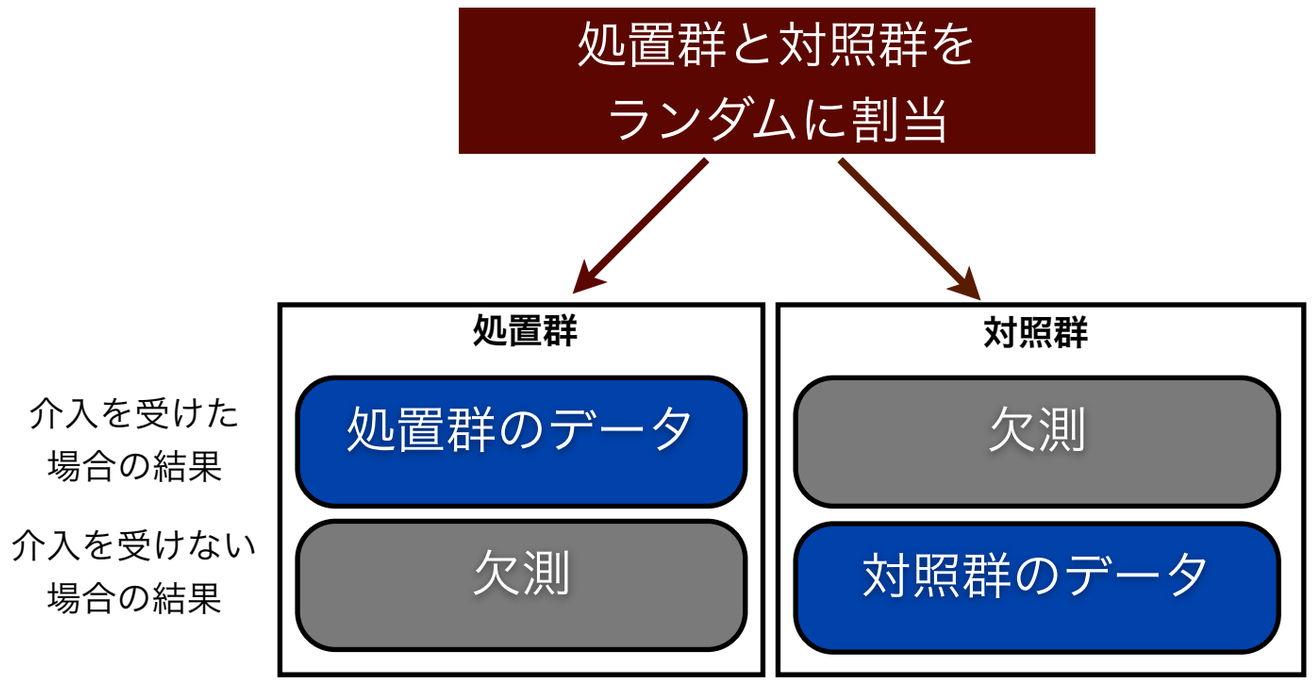
因果効果は単純な処置群と対照群の差になる



因果効果 = 処置群の平均 - 対照群の平均

実験出来るデータ

因果効果は単純な処置群と対照群の差になる



因果効果 = 処置群の平均 - 対照群の平均

実験出来るデータ

因果効果は単純な処置群と対照群の差になる

処置群と対照群を
ランダムに割当

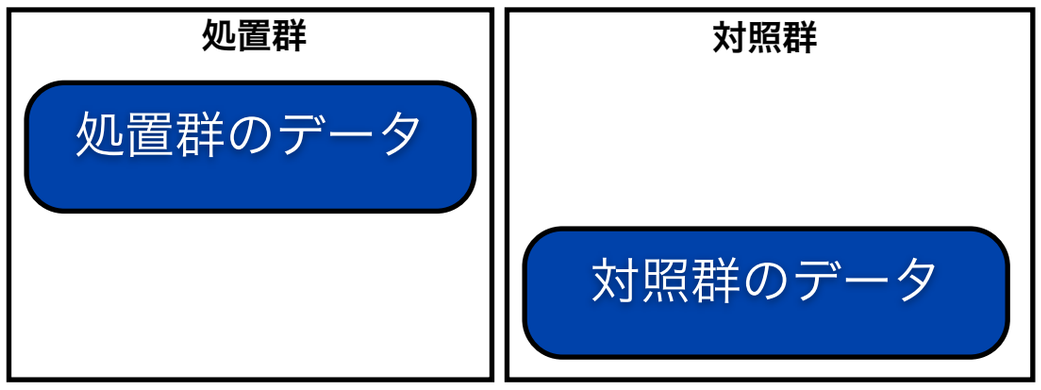
	処置群	対照群	
介入を受けた場合の結果	処置群のデータ	欠測	・対照群が介入を受けていた場合の期待値と介入を受けた処置群の期待値は同じ
介入を受けない場合の結果	欠測	対照群のデータ	・処置群が介入を受けない場合の期待値と介入を受けなかった対照群の期待値は同じ

因果効果 = 処置群の平均 - 対照群の平均

実験出来ないデータ

割当によって処置群と対照群に差が生じるため

単純に比較することが出来ない

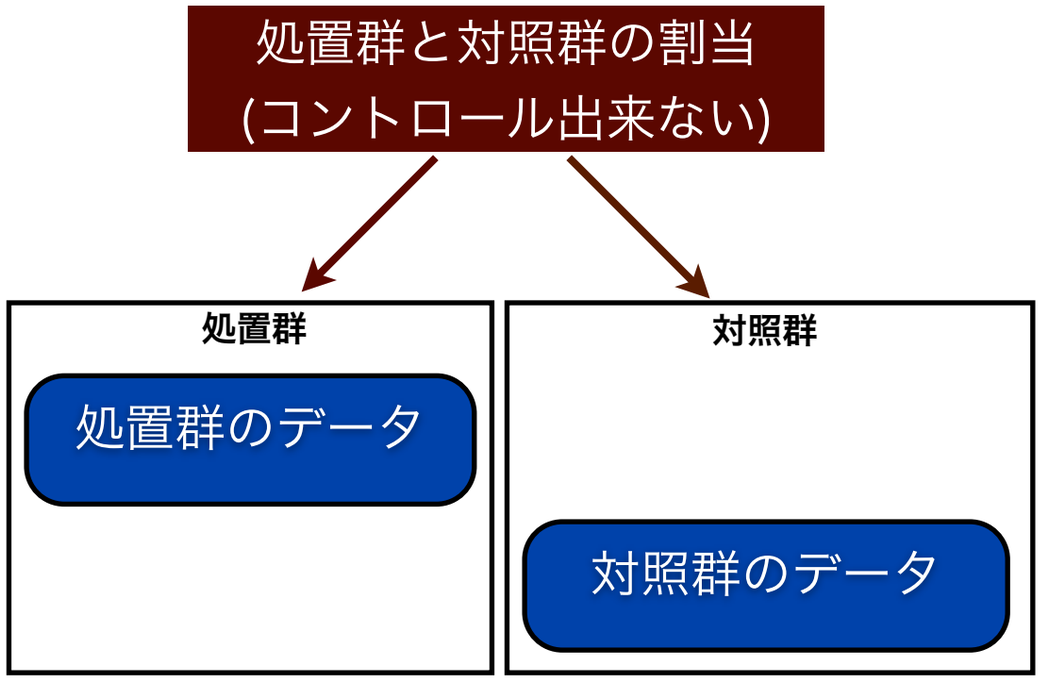


因果効果 ≠ 処置群の平均 - 対照群の平均

実験出来ないデータ

割当によって処置群と対照群に差が生じるため

単純に比較することが出来ない

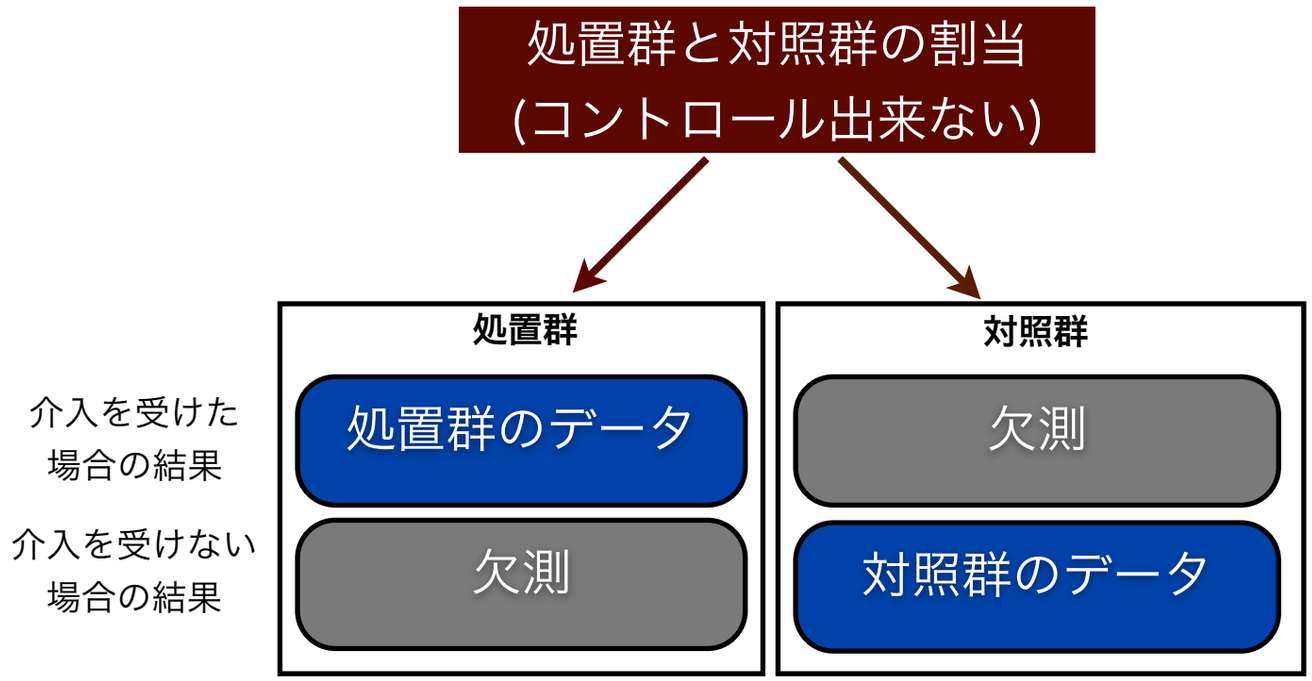


因果効果 ≠ 処置群の平均 - 対照群の平均

実験出来ないデータ

割当によって処置群と対照群に差が生じるため

単純に比較することが出来ない



因果効果 ≠ 処置群の平均 - 対照群の平均

実験出来ないデータ

割当によって処置群と対照群に差が生じるため

単純に比較することが出来ない

処置群と対照群の割当
(コントロール出来ない)

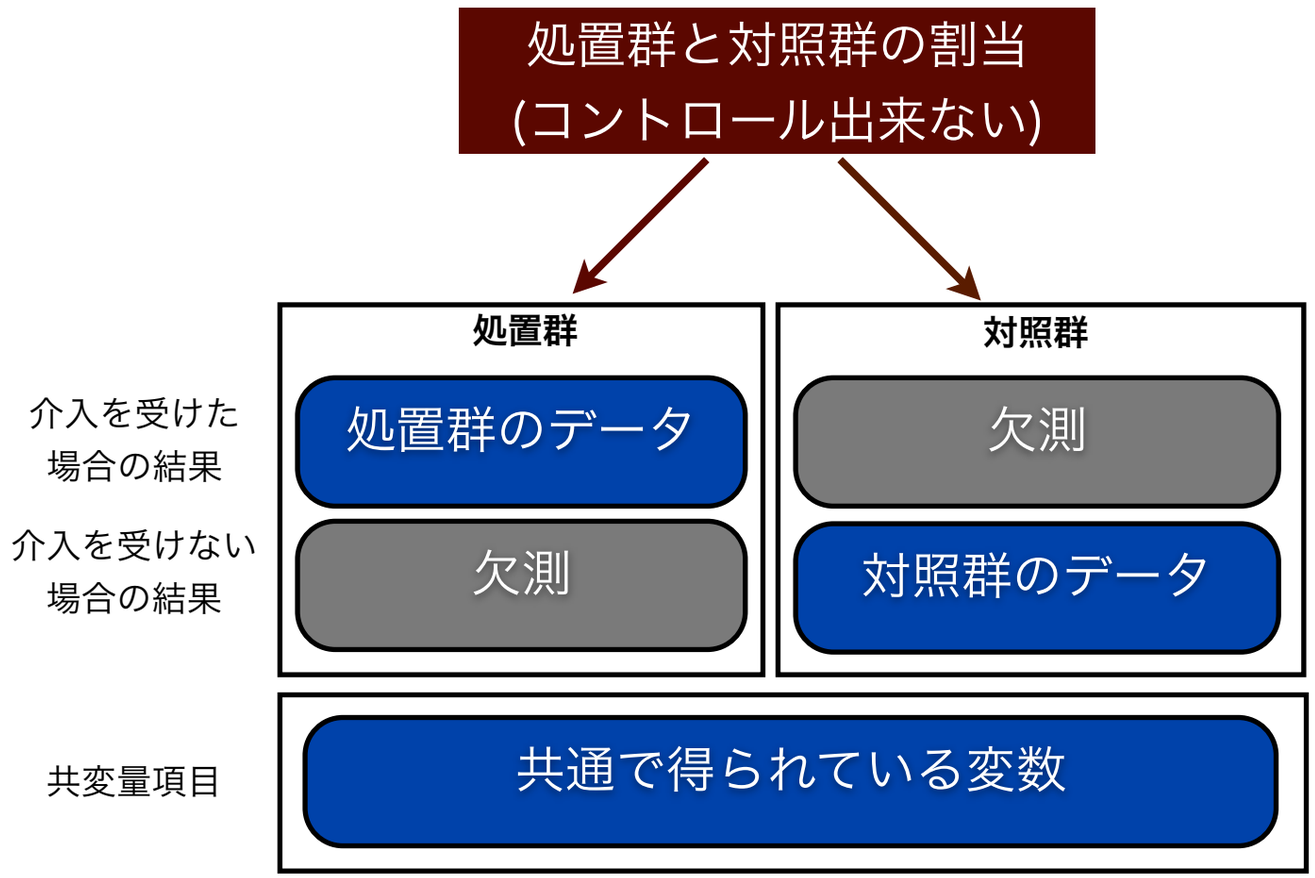
	処置群	対照群	
介入を受けた場合の結果	処置群のデータ	欠測	・対照群が介入を受けていた場合の期待値と介入を受けた処置群の期待値が異なる
介入を受けない場合の結果	欠測	対照群のデータ	・処置群が介入を受けない場合の期待値と介入を受けなかった対照群の期待値が異なる

因果効果 ≠ 処置群の平均 - 対照群の平均

共変量調整

割当や結果変数に影響している共通の変数を用いて

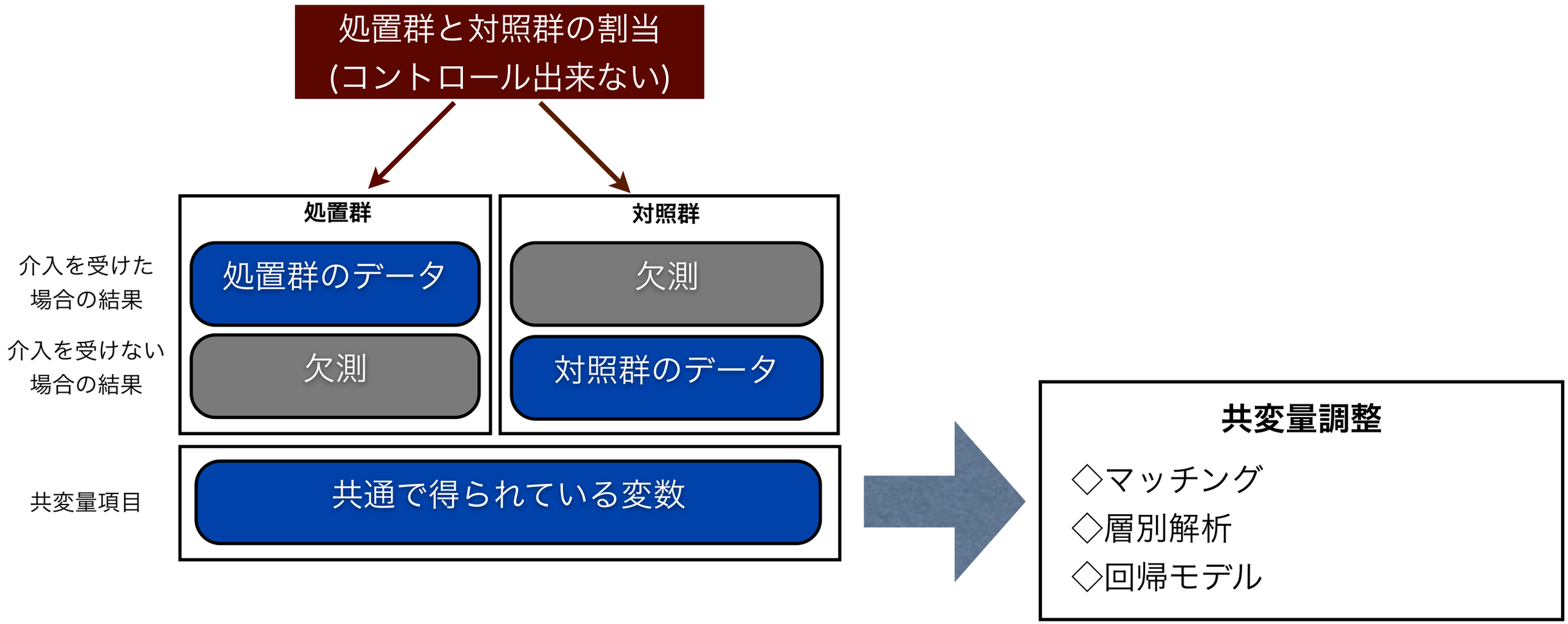
因果効果以外の効果を除去する



共変量調整

割当や結果変数に影響している共通の変数を用いて

因果効果以外の効果を除去する



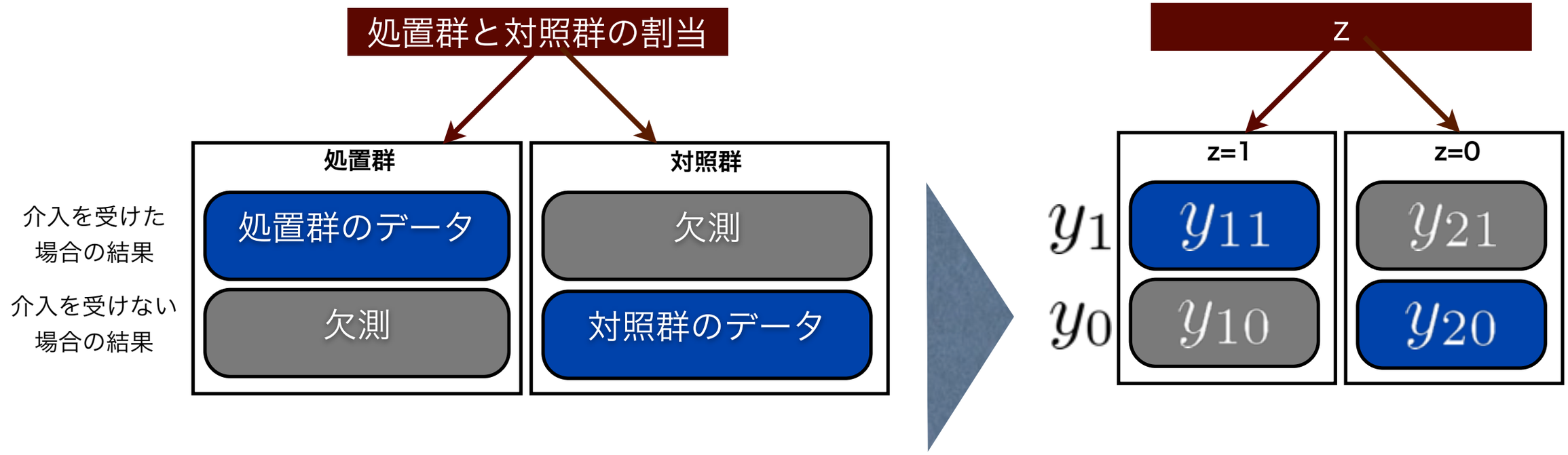
欠測モデル

潜在的結果変数を考える

$$y = zy_1 + (1 - z)y_0$$

欠測モデル

割当変数zと求めたい因果効果

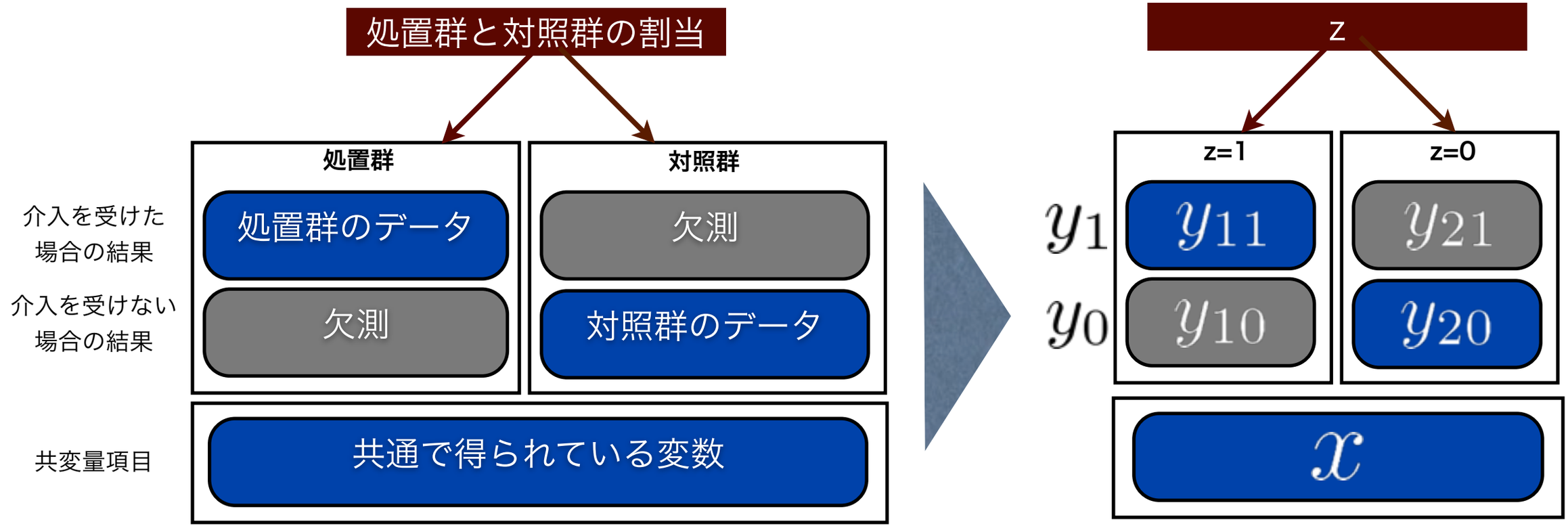


$$TET = E(y_1 - y_0 | z = 1)$$

処置群での平均介入効果
average treatment
effect on the treated

欠測モデル

共変量の影響を除去した因果効果



強く無視出来る割当条件：$(y_1, y_0) \perp\!\!\!\perp z | \boldsymbol{x}$

共変量の影響を除去した因果効果：$E(y_1 - y_0|\boldsymbol{x}) = E(y_1|z=1,\boldsymbol{x}) - E(y_0|z=0,\boldsymbol{x})$

共変量調整

- マッチング
  - 処置群と対照群で共変量が同じになる対象者のペアを作り差をとり、ペア数分の平均を取る
- 層別解析
  - 共変量の値をいくつかの層に分け、層ごとで2つのグループがその共変量の値について等質になるようにし、比較した結果を結合する
- 回帰モデル
  - 各群ごとに回帰関数E(y1|z=1,x)とE(y0|z=0,x)をデータから推定し、その差の標本平均を取る事で因果効果を推定する

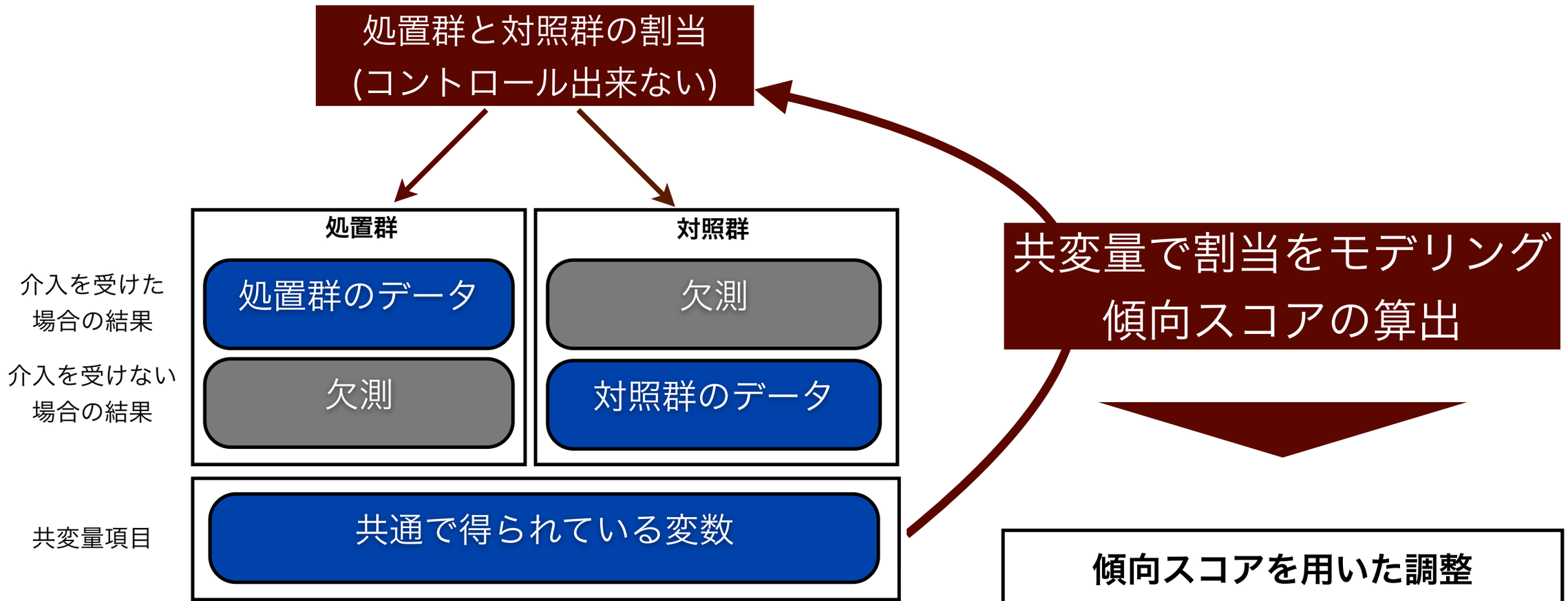
共変量調整の問題点

- マッチング･層別解析での問題
  - 共変量に連続変数があると完全一致のペアは作れない

    次元問題
  - サポート問題
- 回帰モデルでの問題点
  - 結果変数と共変量のモデリングが必要
  - 直接因果効果の推定値は得られない

傾向スコア解析

実験出来ないデータの因果関係を解析する



傾向スコアを用いた調整

◇マッチング
◇層別解析
◇共分散分析

傾向スコアとは

対象者の群1へ割り当てられる確率

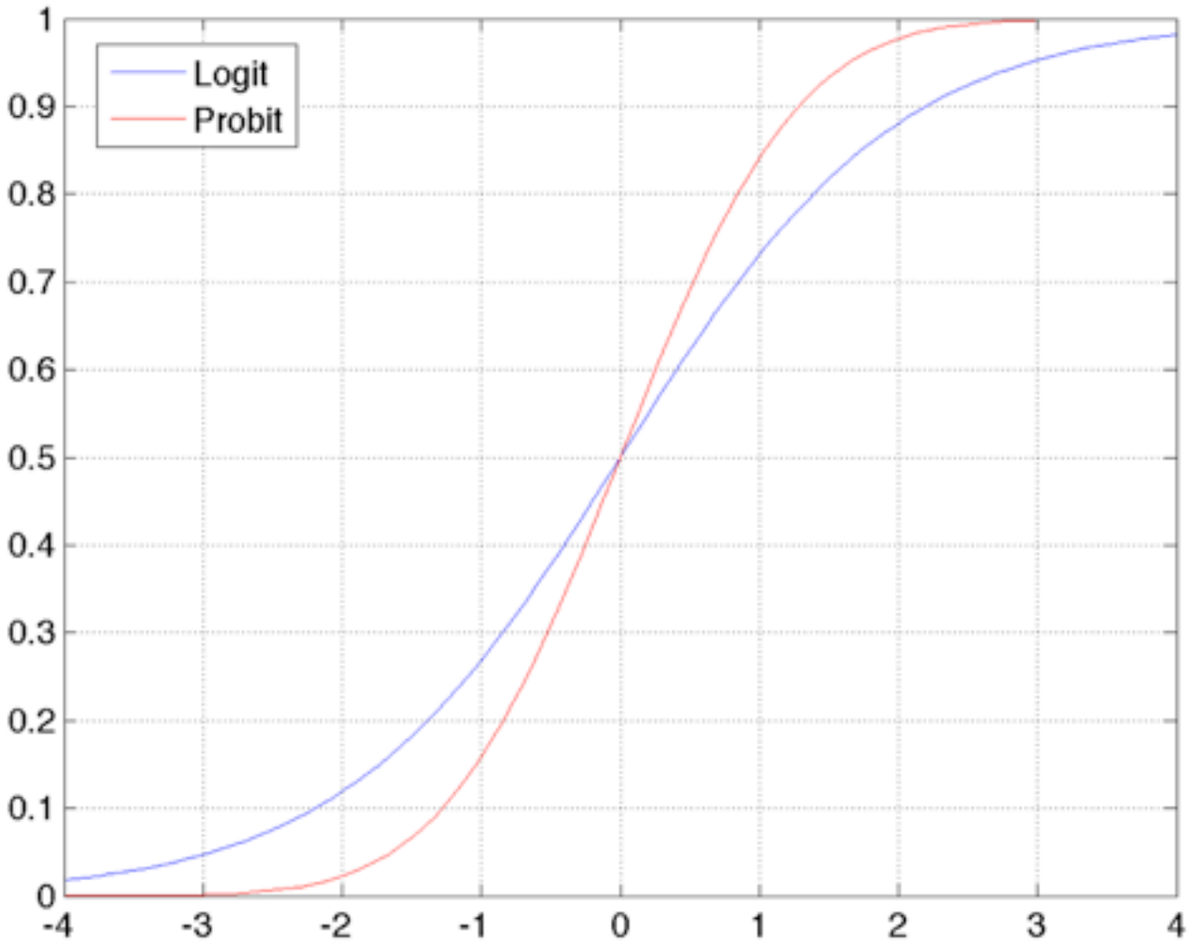
$$e_i = p(z_i = 1 | \boldsymbol{x}_i)$$

z_i ：第i対象者の割当変数の値

$\boldsymbol{x}_i$ ：第i対象者の共変量の値

傾向スコアの推定

プロビット回帰やロジスティック回帰で推定する



$$\hat{e}_i = \int \frac{1}{\sqrt{2\pi}} \exp\{-\frac{z^2}{2}\} dz$$

$$\hat{e}_i = \frac{1}{1 + \exp\{-\hat{\boldsymbol{\alpha}}^t \boldsymbol{x}_i\}}$$

傾向スコアを用いた調整

- マッチング
  - 2つの群で傾向スコアが等しい対象者をペアにしてその差の平均を因果効果とする
- 層別解析
  - 傾向スコアの大小によっていくつかのサブクラスに分け、その各クラスで処置群と対照群の平均の計算と、全体としての効果の推定量を計算する
- 共分散分析
  - 割当変数と傾向スコアを説明変数とした線形の回帰分析を行う

AGENDA

- 自己紹介
- 傾向スコア解析
  - 実験出来るデータ
  - 実験出来ないデータ
  - 共変量調整
  - 傾向スコア推定
  - 傾向スコアを用いた調整
- **Rによる実行**
- 最後に

Rによる実行

Matching パッケージ

Match(Y=NULL, Tr, X, caliper=F,...)

Y ：結果ベクトル

Tr : 割当ベクトル

X ：共変量または傾向スコア

caliper : キャリパーマッチングをやる場合にTRUE

...

Rでマッチング実行

```
install.packages("Matching")
library(Matching)

##
## Matching
##

data(lalonde)

Y78 <- lalonde$re78
Tre <- lalonde$treat
logi <- glm(treat~., data=lalonde[,-9],family=binomial)

## default
summary(Match(Y=Y78, Tr=Tre,X=logi$fitted))

##
summary(Match(Y=Y78, Tr=Tre,X=logi$fitted,M=2))

## caliper matching
summary(Match(Y=Y78, Tr=Tre,X=logi$fitted,caliper=T))
```

Rでカーネルマッチング実行

```
23  ##
24  ## kernel matching
25  ##
26
27  kmy <- lalonde$re74
28  ivec1 <- lalonde$treat
29  estp <- logi$fitted
30  km <- cbind(kmy,estp, ivec1)
31  km1 <- subset(km, ivec1==1)
32  km2 <- subset(km, ivec1==0)
33  km1x <- km1[,2]
34  km1y <- km1[,1]
35  km2x <- km2[,2]
36  km2y <- km2[,1]
37  bw1 <- 1.06*(nrow(km1))^(-0.2) * sd(km1x)
38  bw2 <- 1.06*(nrow(km2))^(-0.2) * sd(km2x)
39  esty1 <- ksmooth(x=km1x,y=km1y,kernel="normal",
40                   bandwidth=bw1,x.points=km2x)
41  esty0 <- ksmooth(x=km2x,y=km2y,kernel="normal",
42                   bandwidth=bw2,x.points=km1x)
43
44  head(esty1$y)
45
46  head(esty0$y)
47
```

AGENDA

- 自己紹介
- 傾向スコア解析
  - 実験出来るデータ
  - 実験出来ないデータ
  - 共変量調整
  - 傾向スコア推定
  - 傾向スコアを用いた調整
- Rによる実行
- **最後に**

次回以降の

発表者・LTを募集しています！

ご清聴ありがとうございました

AGENDA

- 自己紹介
- 傾向スコア解析
  - 実験出来るデータ
  - 実験出来ないデータ
  - 共変量調整
  - 傾向スコア推定
  - 傾向スコアを用いた調整
- Rによる実行
- 最後に

参考文献

- 調査観察データの統計科学
- Package ‘Matching’