2022年3月27日
第24回春の合宿セミナー（日本行動計量学会）
（統計的因果推論入門）

この講義資料は，時間があまったら解説します

講義8b
回帰不連続デザインの発展的事項

長崎大学 情報データ科学部 准教授
高橋 将宜
博士（理工学）
m-takahashi@nagasaki-u.ac.jp

概要

- カーネル密度推定
- バンド幅の選択
- RDプロット
- 連続性の仮定と強制変数の操作
- 共変量の活用

高橋（2022, pp.232-254）

カーネル密度推定

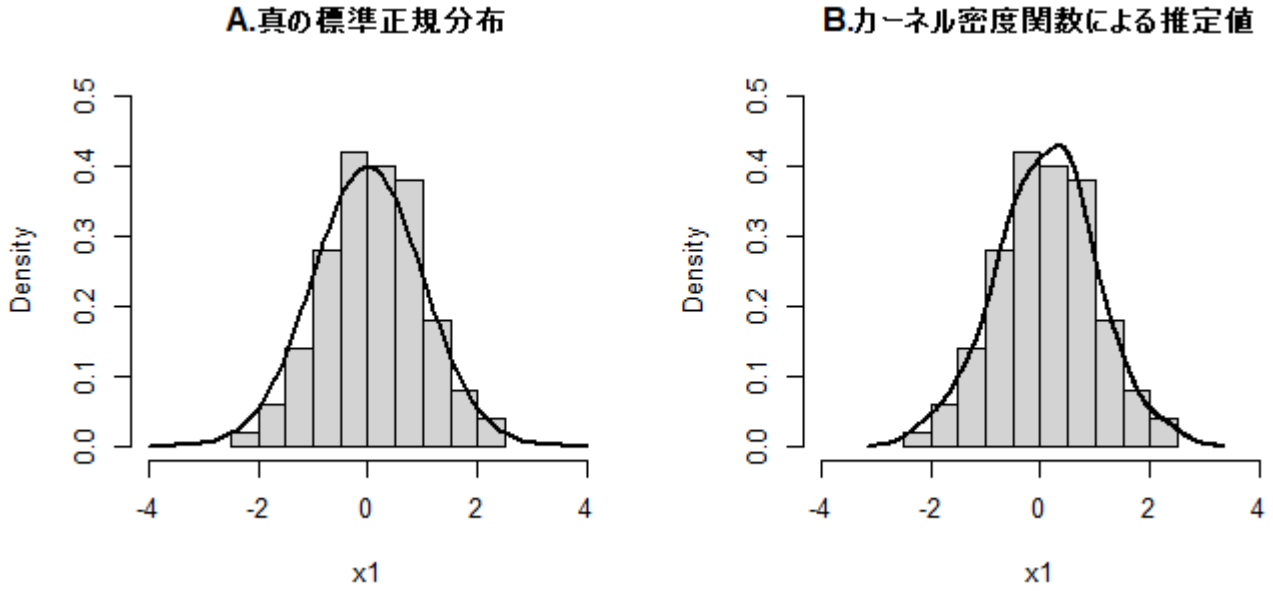
カーネル密度推定値

- R関数rdrobust
  - kernelの引数をuniformとしていた.
  - これが何を意味していたのか確認しよう.

- カーネル密度推定値
  - ヒストグラムをスムーズ化したもの
  - 観察されたデータから確率密度関数を推定する方法

図16.5

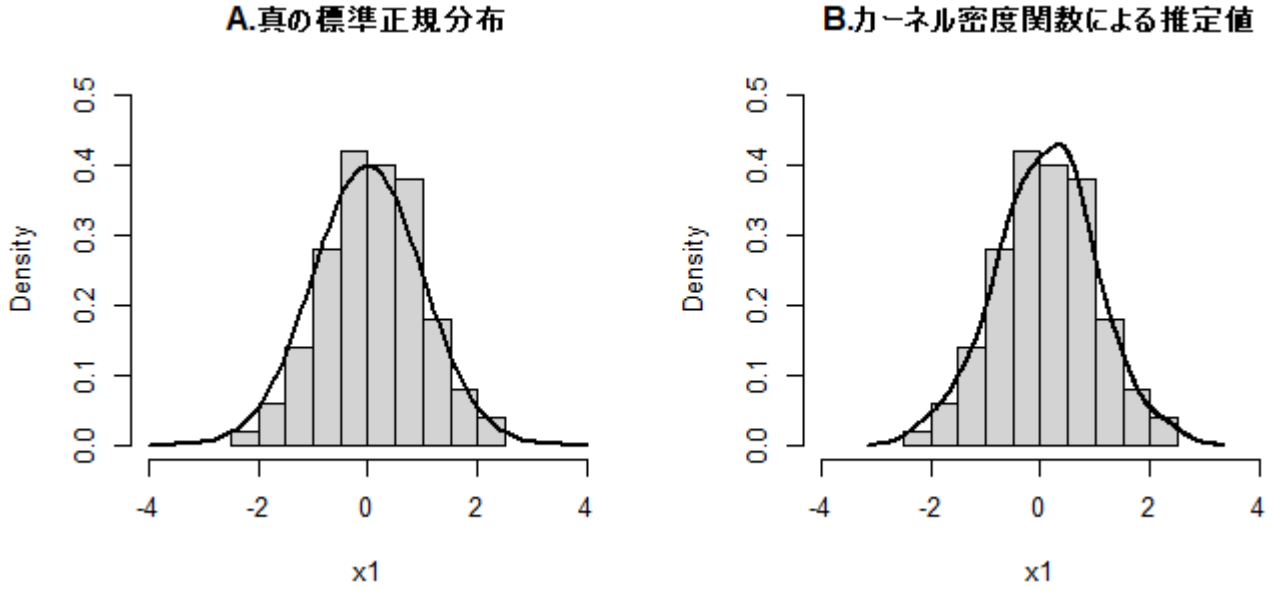
高橋（2022, p.232）



- 図16.5A：凹凸のあるヒストグラムに真の標準正規分布を重ねて表示している.
- 図16.5B：凹凸のあるヒストグラムをスムーズ化した曲線を表示している. これがカーネル密度推定値である.
- ヒストグラムのバーの幅（bin）：一定で固定

カーネル密度推定値の計算方法

高橋（2022, p.232）



- ヒストグラムのバーのようなウィンドウを左から右に少しずつ動かしていき，あるxの値に対して，以下の式によって密度推定値を計算する．

$$\hat{p}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

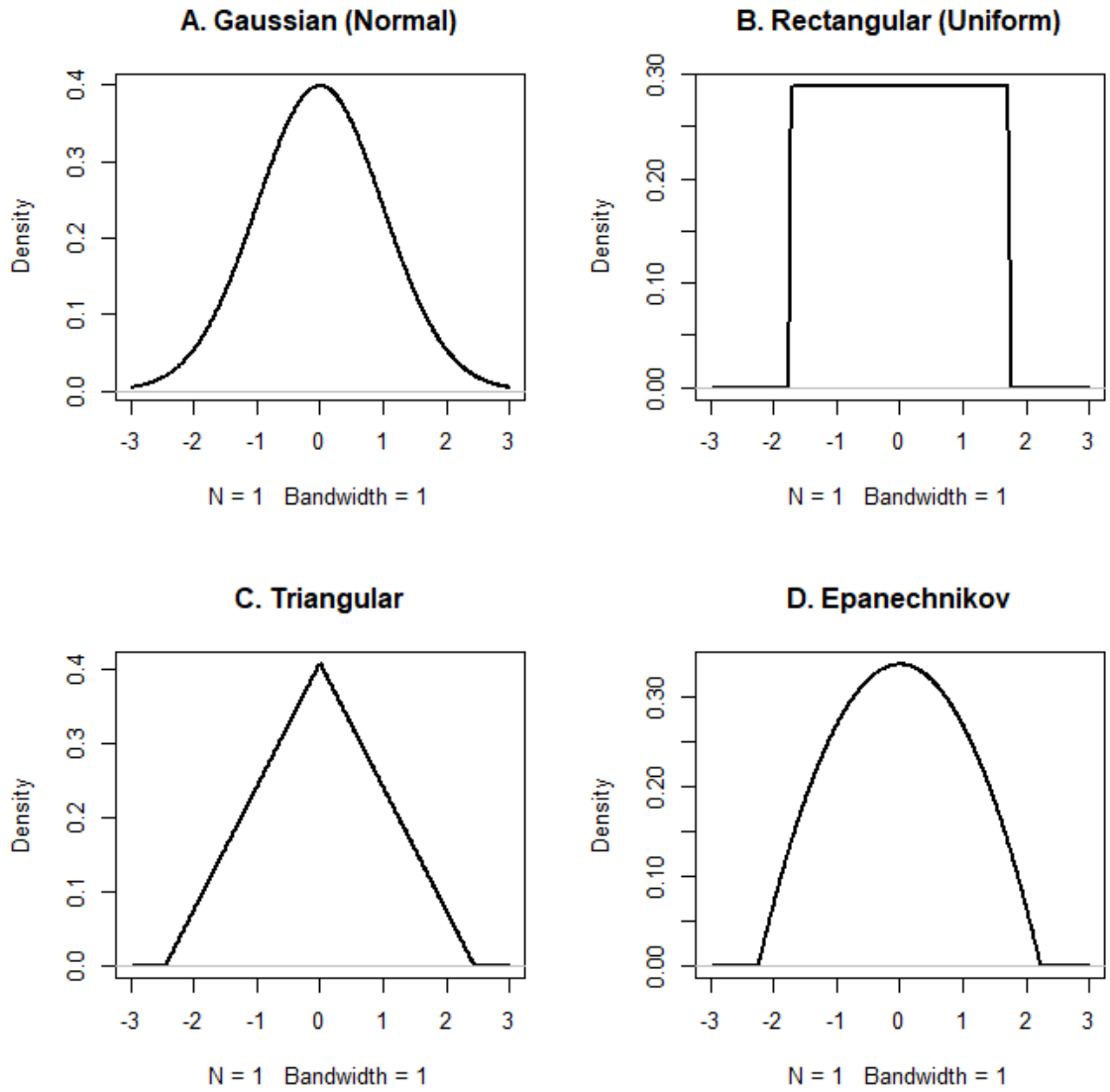
  - x_iはn個のデータの値であり，$K(\cdot)$は最頻値が0で面積が1の対称な密度関数，hはウィンドウの幅（バンド幅）

$K(\cdot)$：カーネル関数

- ガウス関数
  - Gaussian function
  - $N(0,1)$
- 矩形関数
  - rectangular functionまたはuniform function
  - $U(-a,a)$
- 三角形関数
  - triangular function
  - $1-|t|$
- エパネチニコフ関数
  - Epanechnikov function
  - $[3(1-t^2)]/4$

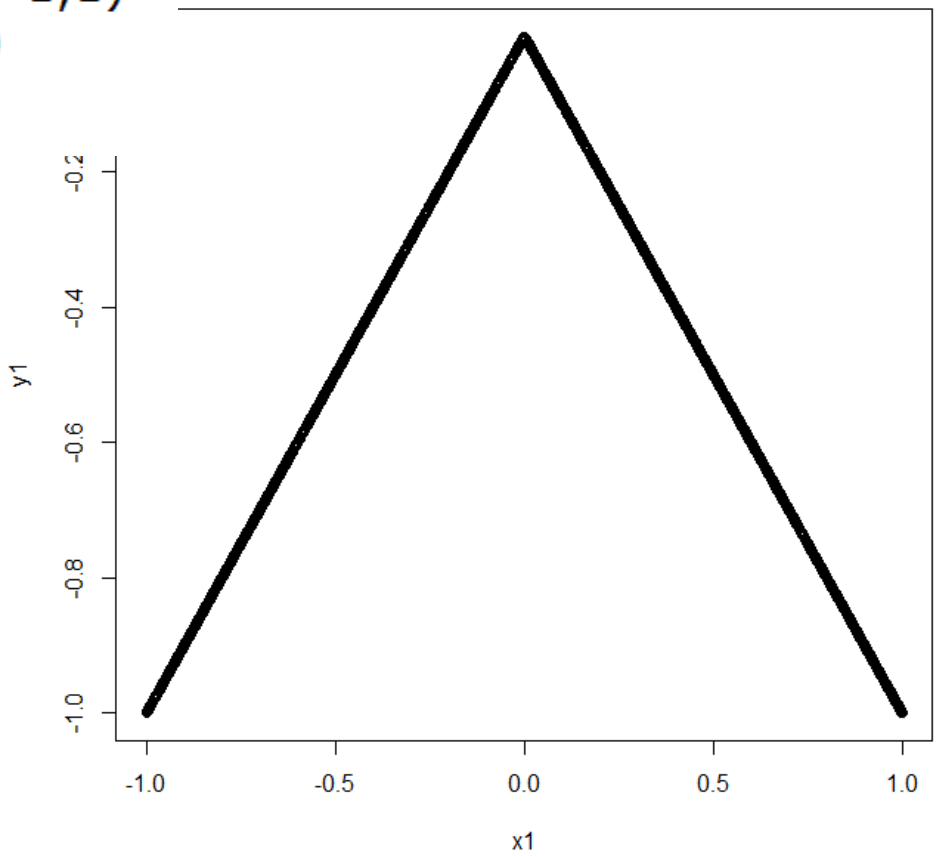
カーネル関数の形状

高橋（2022, p.233）



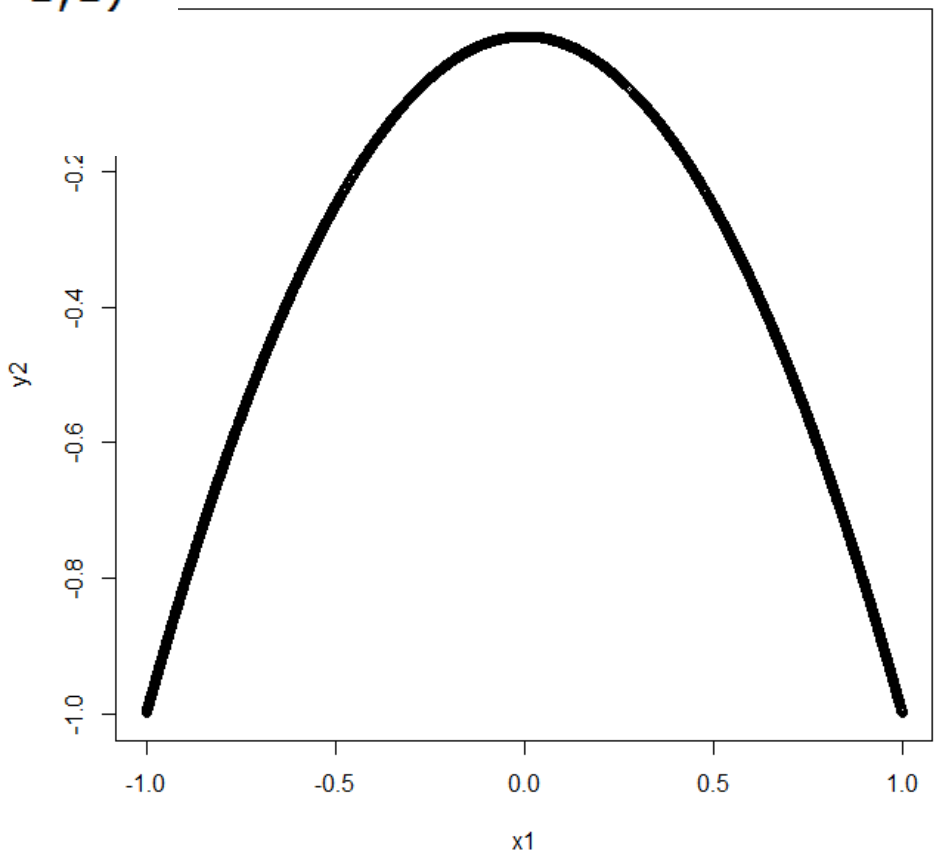
三角形関数：$y = -|x|$

```
n1 <- 10000
x1 <- runif(n1,-1,1)
y1 <- -abs(x1)
plot(x1,y1)
```



エパネチニコフ関数：$y = -x^2$

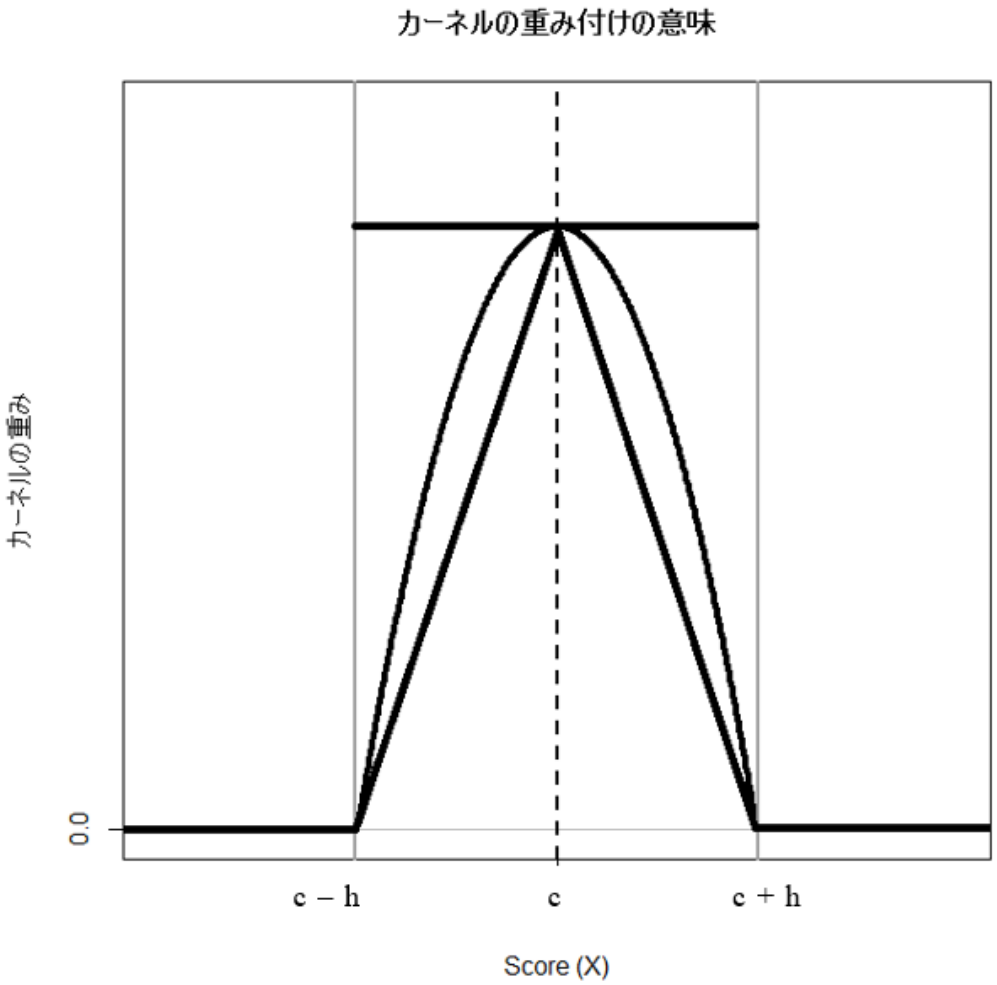
```
n1 <- 10000
x1 <- runif(n1,-1,1)
y2<- -x1^2
plot(x1,y2)
```



カーネル密度推定

図16.9

高橋（2022, p.236）



バンド幅の選択

バンド幅の広さ

- 狭いバンド幅
  - 局所的な無作為割付けの成立している可能性が高まるため推定値の偏りは小さくなる
  - 使用できる観測数が減るため精度が低くなる
- 広いバンド幅
  - 推定値の精度は上がる
  - 偏りが大きくなる
    - あまりにもバンド幅が大きすぎると，もはや閾値の周辺の値を比較しているとはいえず，回帰不連続デザインの意義がなくなってしまう.
- 偏りとばらつきの最適なバランスを探す問題

平均二乗誤差（MSE: mean squared error）

- 偏りとばらつきの大きさをバランスよく評価する指標

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

  - 推定量$\hat{\theta}$の分散に偏りの二乗を加えたものに変形できるので，偏りとばらつきのバランスを取った指標
  - 平均二乗誤差（MSE）の最も小さな推定量が，最も良い推定量と見なす考え方

IKバンド幅（IK bandwidth）

高橋（2022, p.237）

$$\hat{h}_{opt} = C_K \left(\frac{\hat{\sigma}_-^2(c) + \hat{\sigma}_+^2(c)}{\hat{f}(c)\left(\hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c)\right)^2 + \hat{r}_- + \hat{r}_+} \right)^{1/5} N^{-1/5}$$

- MSEの意味で最適なバンド幅
  - 出典：Imbens and Kalyanaraman（2012）

data15

- data15 <- read.csv(file.choose( ))
- attach(data15)
- summary(data15)

```
> summary(data15)
      y0t              y1t              y3              t1              x1       
 Min.   : 65.31   Min.   : 53.89   Min.   : 65.31   Min.   :0.000   Min.   : 62.55  
 1st Qu.:132.72   1st Qu.:112.83   1st Qu.:115.39   1st Qu.:1.000   1st Qu.:133.61  
 Median :150.91   Median :122.92   Median :124.41   Median :1.000   Median :150.56  
 Mean   :151.17   Mean   :122.19   Mean   :123.92   Mean   :0.798   Mean   :150.44  
 3rd Qu.:169.47   3rd Qu.:131.75   3rd Qu.:132.81   3rd Qu.:1.000   3rd Qu.:167.29  
 Max.   :229.33   Max.   :174.08   Max.   :174.08   Max.   :1.000   Max.   :235.05  
```

Rパッケージrdrobustのrdbwselect_2014関数

- library(rdrobust)
- IKband <- rdbwselect_2014(y3, x1, c=130, bwselect="IK")
- IKband$bws

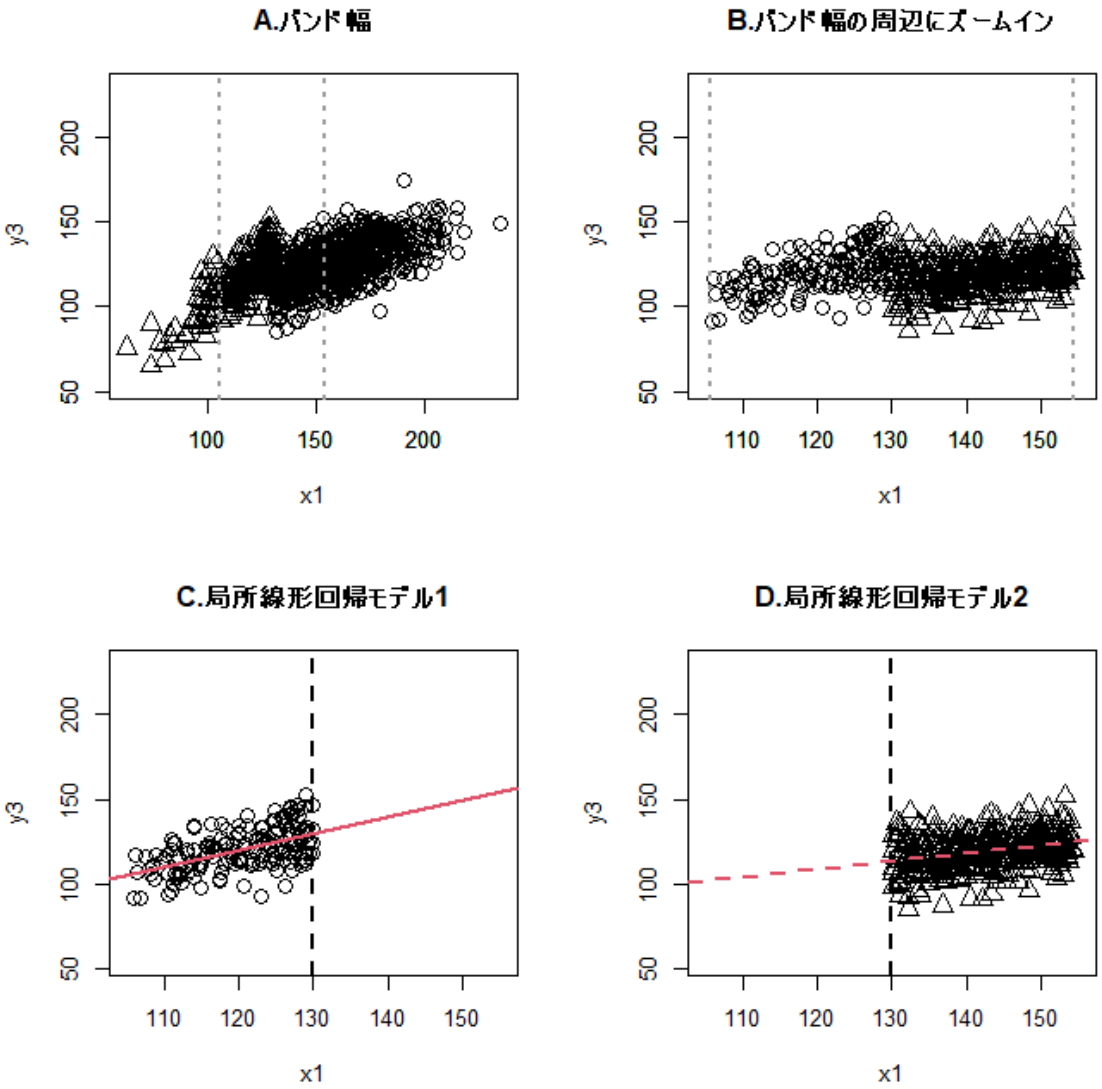
```
> library(rdrobust)
> IKband <- rdbwselect_2014(y3, x1, c=130, bwselect="IK")
> IKband$bws
             h        b
[1,] 24.3555 19.35648
```

R関数rdrobustによる回帰不連続デザイン（1）

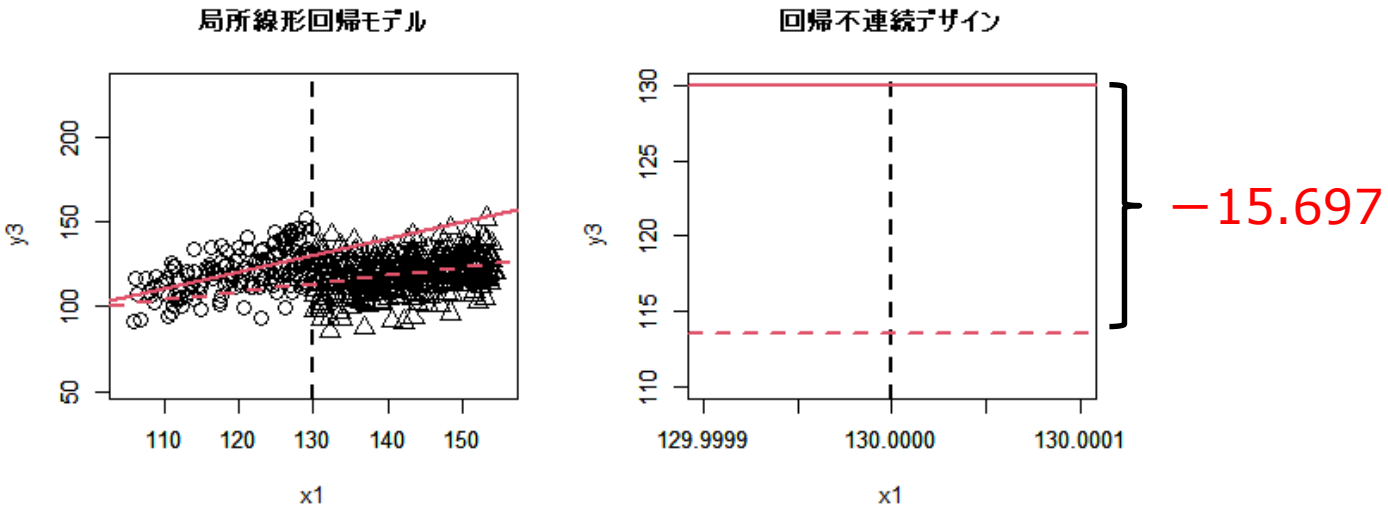
- model3 <- rdrobust(y3, x1, c=130, h=24.3555)
- summary(model3)

```
Call: rdrobust

Number of Obs.                 1000
BW type                      Manual
Kernel                   Triangular
VCE method                       NN

Number of Obs.                 202         798
Eff. Number of Obs.            159         349
Order est. (p)                   1           1
Order bias  (q)                  2           2
BW est. (h)                 24.355      24.355
BW bias (b)                 24.355      24.355
rho (h/b)                    1.000       1.000
Unique Obs.                    197         749

=============================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
=============================================================================
  Conventional   -15.697     2.291    -6.851     0.000   [-20.189 , -11.206]
        Robust         -         -    -4.220     0.000   [-20.544 , -7.513]
=============================================================================
```

R関数rdrobustによる回帰不連続デザイン（1）の図解a



R関数rdrobustによる回帰不連続デザイン（1）の図解b

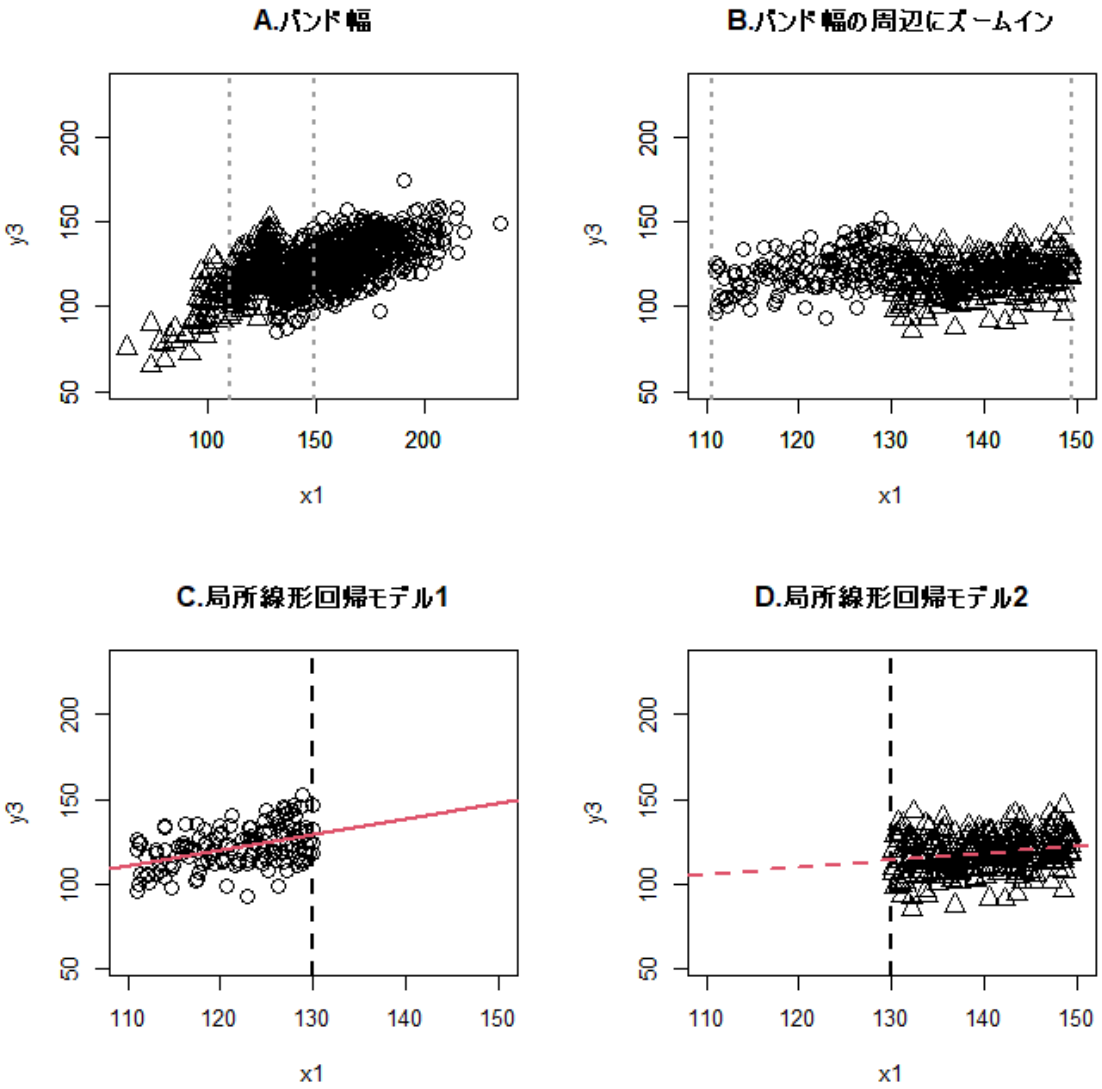


R関数rdrobustによる回帰不連続デザイン（2）

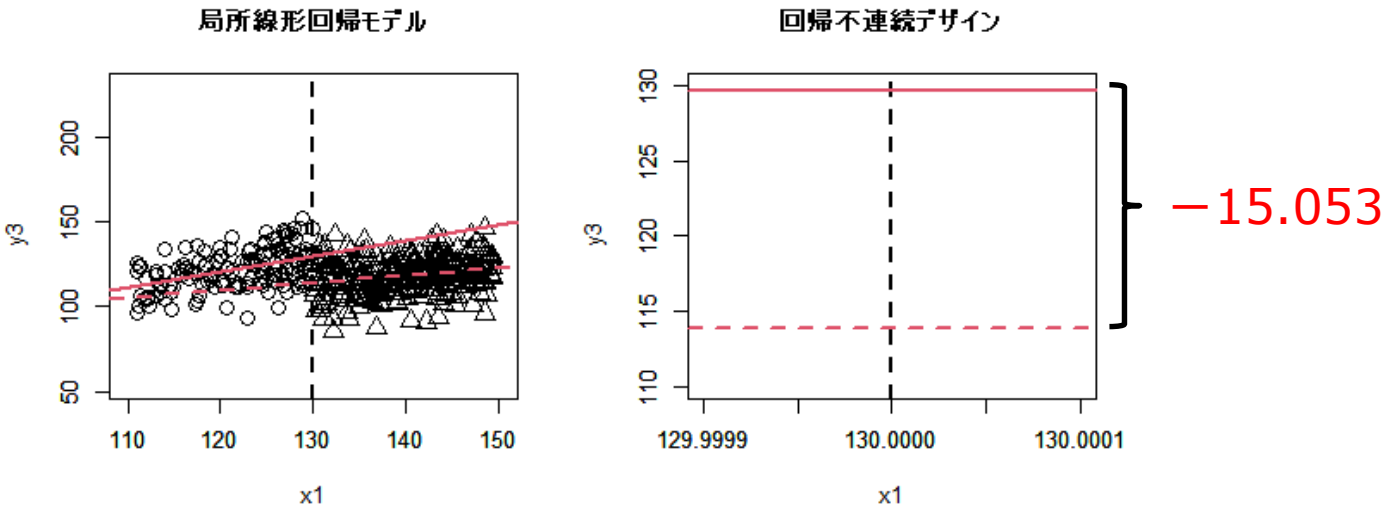
- model4 <- rdrobust(y3, x1, c=130, h=19.35648)
- summary(model4)

```
Call: rdrobust

Number of Obs.                 1000
BW type                      Manual
Kernel                   Triangular
VCE method                       NN

Number of Obs.                 202         798
Eff. Number of Obs.            144         274
Order est. (p)                   1           1
Order bias  (q)                  2           2
BW est. (h)                 19.356      19.356
BW bias (b)                 19.356      19.356
rho (h/b)                    1.000       1.000
Unique Obs.                    197         749

=============================================================================
        Method     Coef. Std. Err.          z     P>|z|      [ 95% C.I. ]
=============================================================================
  Conventional   -15.053     2.558     -5.884     0.000   [-20.067 , -10.038]
        Robust         -         -     -3.777     0.000   [-20.996 , -6.649]
=============================================================================
```

R関数rdrobustによる回帰不連続デザイン（2）の図解a



R関数rdrobustによる回帰不連続デザイン（2）の図解b



カバー率の誤差の意味で最適なバンド幅

- Calonico et al.（2018, pp.767-768）
  - MSEの意味で最適なバンド幅を使った場合，信頼区間の大きさが不適切になることを指摘
  - カバー率の誤差（coverage error）の意味で最適なバンド幅を用いることで，偏りを是正し，適切な信頼区間を構築できると提案
  - 引数bwselectにおいてcerrdと指定
  - coverage error rate regression discontinuity

R関数rdrobustによる回帰不連続デザイン（3）

- model5 <- rdrobust(y3, x1, c=130, bwselect="cerrd")
- summary(model5)

```
Call: rdrobust

Number of Obs.                 1000
BW type                       cerrd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                 202         798
Eff. Number of Obs.             98         156
Order est. (p)                   1           1
Order bias  (q)                  2           2
BW est. (h)                 11.183      11.183
BW bias (b)                 24.521      24.521
rho (h/b)                    0.456       0.456
Unique Obs.                    197         749

=============================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
=============================================================================
  Conventional    -14.177     3.316    -4.275     0.000   [-20.677 , -7.677]
        Robust          -         -    -3.836     0.000   [-20.909 , -6.769]
=============================================================================
```

結局，どのバンド幅がいいの？

- Imbens and Lemieux（2008, p.633）
  - 1つのバンド幅だけを採用して解析するのではなく，複数のバンド幅を使用するべきと指摘
  - データから最適とされるバンド幅を半分にしたり，2倍にしたり，いくつかのパターンを解析して，バンド幅の選び方によって解析結果がどのように変化するか，検討するべき

	点推定値	標準誤差	95%CI下限	95%CI上限
MSE最適h	-15.697	2.291	-20.189	-11.206
MSE最適b	-15.053	2.558	-20.067	-10.038
CER最適	-14.177	3.316	-20.677	-7.677

Takahashi（2021）

- 多重代入法不連続デザイン（MIRDD: multiple imputation regression discontinuity design）
  - 閾値における局所的な処置効果を可視化して分析でき，異なる大きさのバンド幅を用いた場合に，結果がどのように視覚的に変化するか検証できる.

https://doi.org/10.1080/03610918.2021.1960374

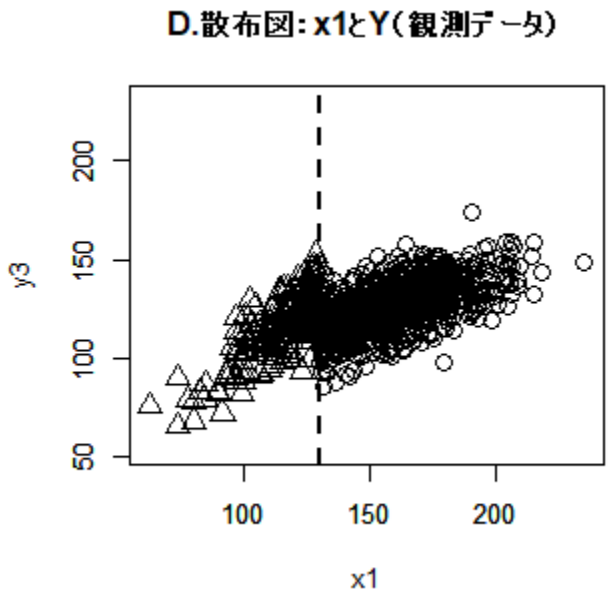
RDプロット

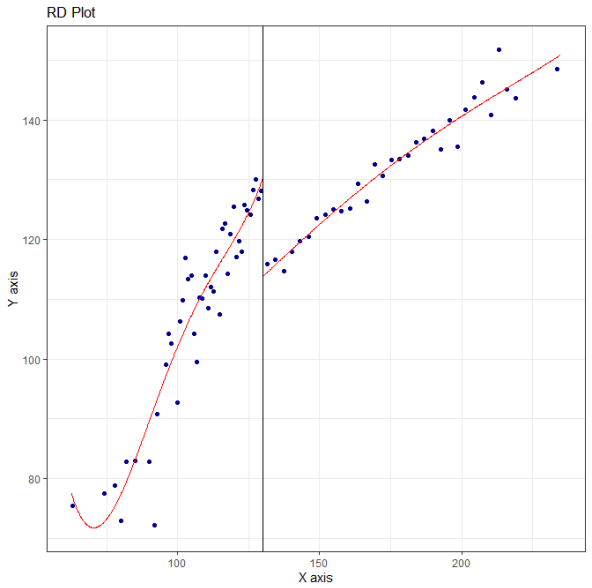
RDプロット

- 散布図の横軸をいくつかのビンに分割して，そのビンの中に入るデータの平均値を図示することで，データをスムーズ化して，その背後にある分布形をあぶり出す
- Rパッケージrdrobust
- rdplot(結果変数, 強制変数, c=閾値)

出力結果（1）

- rdp1 <- rdplot(y3, x1, c=130)
- summary(rdp1)



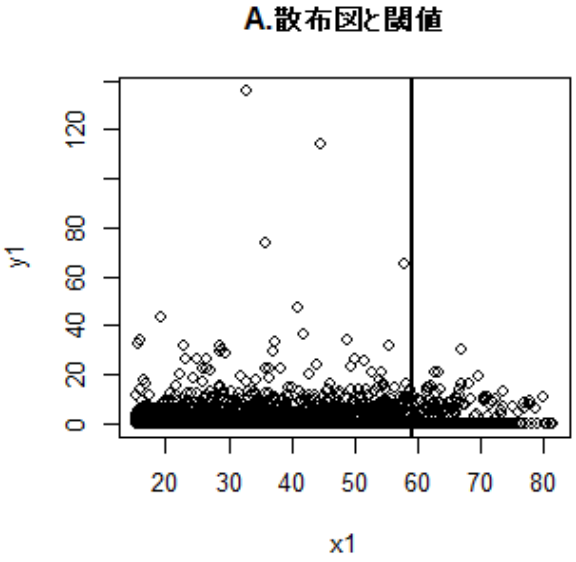


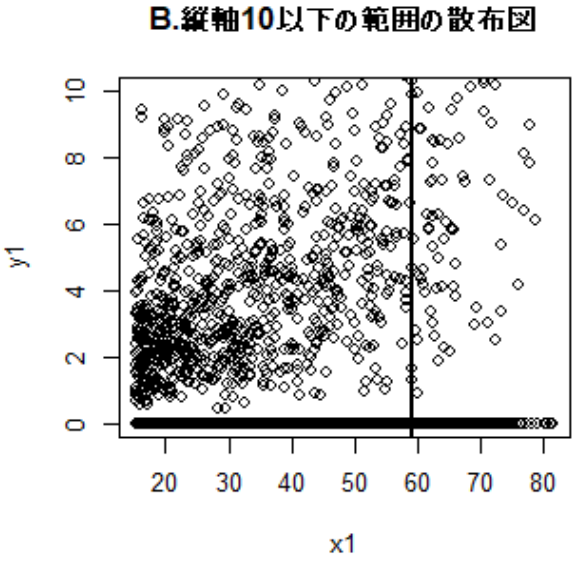
出力結果（2）

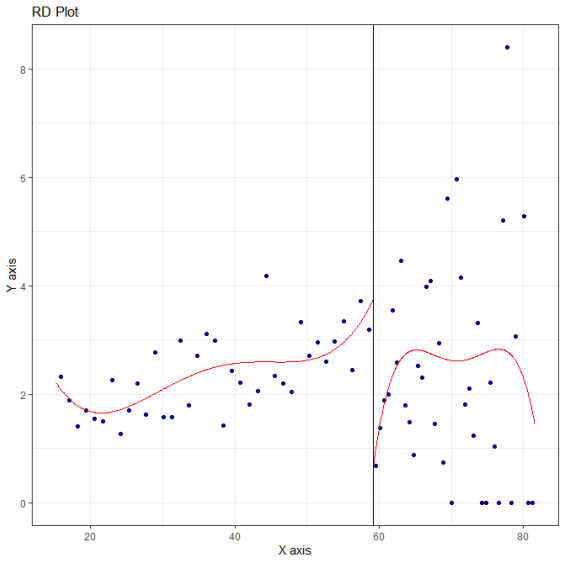
```
> summary(rdpl)
Call: rdplot

Number of Obs.                  1000
Kernel                       Uniform

Number of Obs.                   202          798
Eff. Number of Obs.              202          798
Order poly. fit (p)                4            4
BW poly. fit (h)              67.450      105.050
Number of bins scale               1            1

Bins Selected                     68           36
Average Bin Length             0.992        2.918
Median Bin Length              0.992        2.918

IMSE-optimal bins                 13           14
Mimicking Variance bins           68           36

Relative to IMSE-optimal:
Implied scale                  5.231        2.571
WIMSE variance weight          0.007        0.056
WIMSE bias weight              0.993        0.944
```

図17.3と図17.4

高橋（2022, p.249）







連続性の仮定と強制変数の操作

連続性の仮定が満たされない典型例

- 強制変数の操作（manipulation of the running variable）
  - 閾値の存在が知られていて，強制変数Xの値が操作できる場合
- 具体例
  - 入学試験で90点以上であれば学費が免除され，90点未満ならば学費は免除されないとし，この情報はオープンになっているものとする.
  - 一部の受験生が試験問題を事前に入手している場合
  - 一部の受験生の試験結果に対して採点者が不正に加点している場合
  - 意図的に90点以上を取ることができる

強制変数の操作

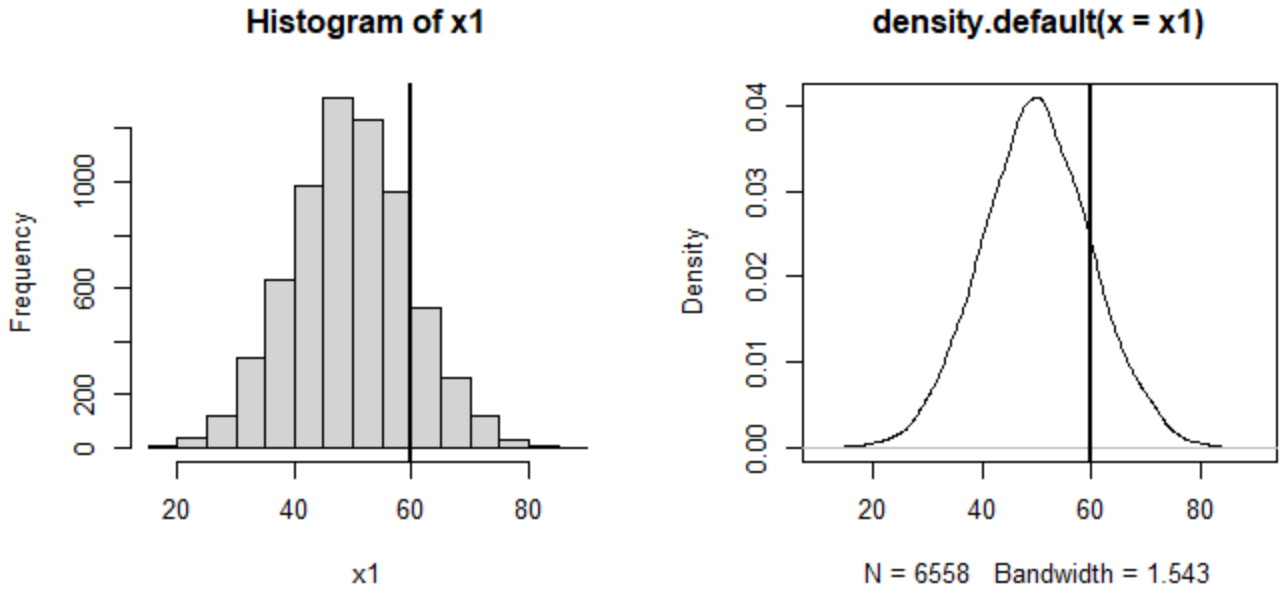
- 閾値がどこにあるかが単に知られているだけでなく，強制変数の値が正確に操作されるという意味
  - 試験結果と授業料免除の例
  - 何も勉強せずに試験を受けに行く人はほとんどいない
  - 学生は実際に強制変数に対して何らかの操作を行っている

夫婦共働き世帯の収入の例

- 夫：フルタイムで働く
- 妻：パートタイムで働く
  - 所得税の所得控除を考えて，パートから得られる年間収入が103万円以下に抑えられる傾向のあることが知られている.
  - 103万円のちょうど上と下の分布を比べると，103万円よりわずかに少ない年収の人の数が，103万円以上の年収の人よりも多くなっているはず

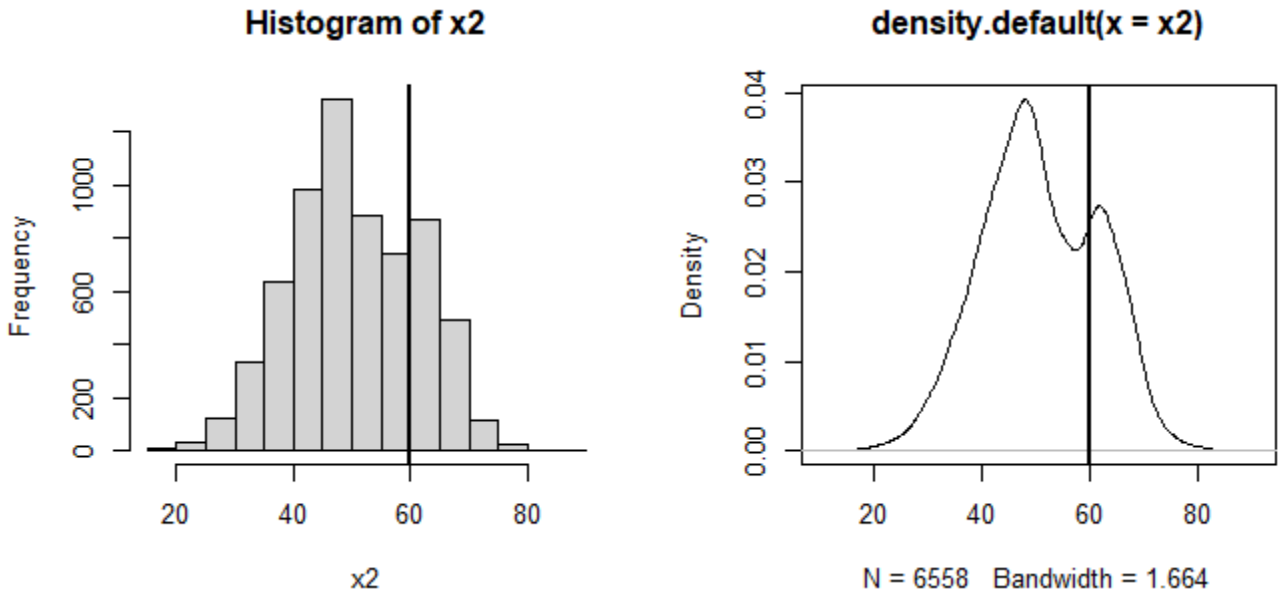
強制変数は操作されていない

高橋（2022, p.244）

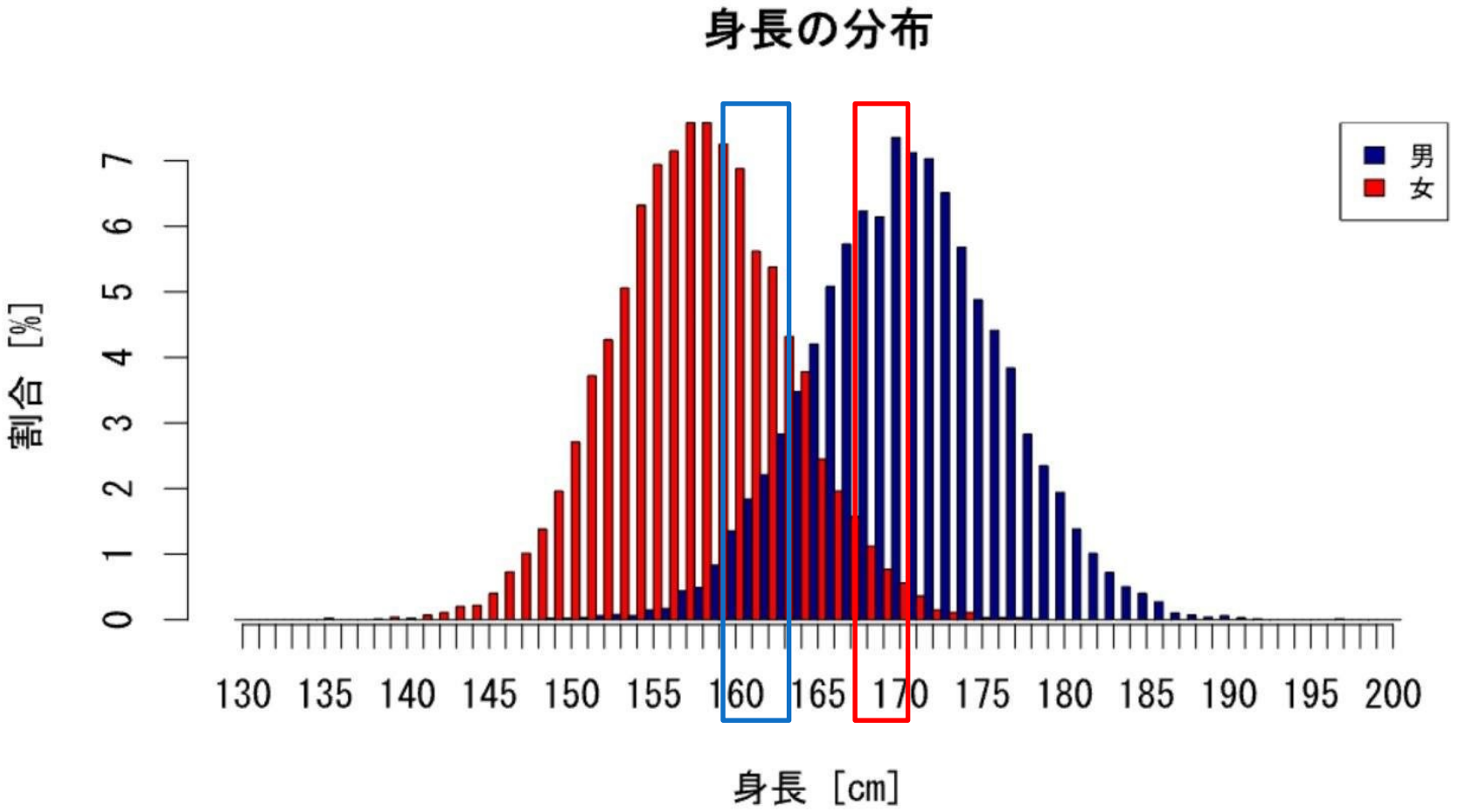


強制変数は操作されている

高橋（2022, p.245）



男女の身長の例



https://twitter.com/justice_woods_/status/1493913335653232641

フォーマルな検定による連続性の仮定の診断

- McCrary（2008）の手法
  - RパッケージrddのDCdensity関数
- Cattaneo et al.（2018a, 2020）の手法
  - 教科書pp.245-246

共変量の活用

回帰不連続デザイン：共変量は必要？

- 回帰不連続デザイン
  - 局所的な無作為割付け
  - 局所的な実験研究
- 共変量をモデルに取り込まなくても交絡を取り除くことができる

実験研究：共変量は必要？

高橋（2022, p.85）

- 実験研究
  - 共分散分析を用いて共変量を活用することで，推定の精度を向上させることができる
  - 教科書p.85, p.89

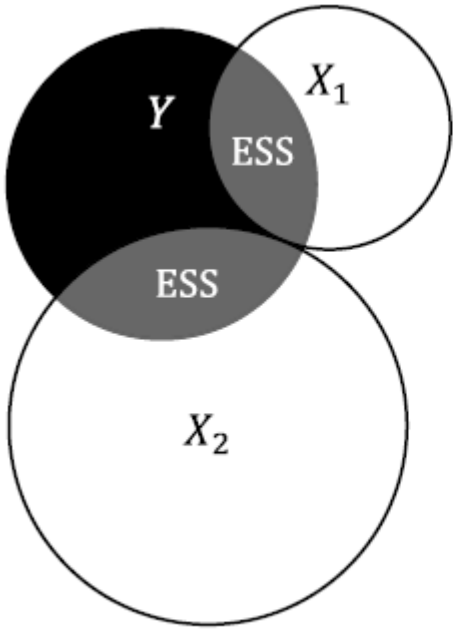


図6.6a

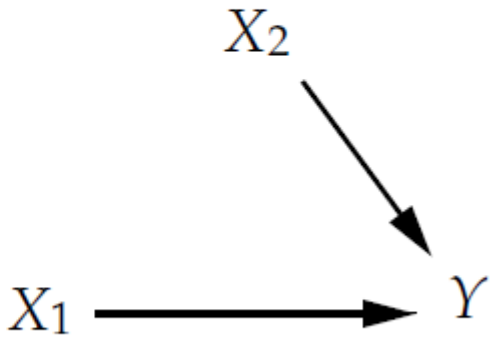


図6.6b

共変量は必須ではないが，あってもよい

- 同じ理屈が回帰不連続デザインにも当てはまる
  - 共変量については関数形を気にする必要はない
  - 単純にモデルに追加すればよい
  - 共変量を追加しても追加しなくても，パラメータ推定値の一致性に影響はないから
    - パラメータの推定値に影響を与えるような共変量をモデルに取り込んではならない
    - もし共変量を取り入れてパラメータの推定値に大きな変化があるとしたら，回帰不連続デザインのモデリング自体がうまく機能していないおそれがある

Calonico et al.（2019）

- 疑問
  - 共変量は解析モデルだけで利用すればよいのか？
  - バンド幅の推定の際にも利用するべきなのか？

- 答え
  - 共変量は，バンド幅の推定の際にも利用した上で，解析モデルにも含めることで，精度が最もよくなる

Rパッケージrdrobustに共変量を追加

- 引数covsの右辺に共変量を指定すればよい
  - 結果変数：y1
  - 強制変数：x1
  - 1個目の共変量：z1
  - 2個目の共変量：z2
  - 閾値：50

```
> rdrobust(y1,x1,c=50,covs=c("z1","z2"))
 as.matrix(covs)[na.ok, , drop = FALSE] でエラー:
   (subscript) 論理値添え字が長すぎます
 追加情報:  警告メッセージ:
 na.ok & complete.cases(covs) で:
   長いオブジェクトの長さが短いオブジェクトの長さの倍数になっていません

> zs<-cbind(z1,z2)
> rdrobust(y1,x1,c=50,covs=zs)
Call: rdrobust
```