#------------------------------------------------------------------

# 昭和大学　第9回実践臨床統計学セミナー　Rハンズオン

# 傾向スコアマッチング

# 2022年2月21日

#------------------------------------------------------------------

# Rのプログラム作成には、高機能なテキストエディタを使用することをお勧めします。

# 初心者の方向けのソフトとしては、Notepad++をお勧めします（フリーソフトです）。以下のURLからダウンロードすることができます。

#　https://github.com/notepad-plus-plus/notepad-plus-plus/releases/download/v7.9.5/npp.7.9.5.Installer.exe

# 拡張子を「.r」として保存したプログラムのファイルを開くと、基本関数やコメント箇所などが色分けされて表示され、効率的なプログラミングを行うことができます。

# Macユーザーの方は、Rのパッケージのインストールができないことがあるようです。

# RStudioを使えば、特に問題なく、インストールができることが多いようですので、もし支障があるようでしたら、そちらをお試しください（以下のURLからインストールパッケージをDLできます）。

# https://download1.rstudio.org/desktop/macos/RStudio-1.4.1717.dmg

#------------------------------------------------------------------

# 1. データセットの読み込み

#------------------------------------------------------------------

setwd("C:\\R2021")

# setwdで、カレントのディレクトリを指定します。

# 他のソフトウェアと違う点は、フォルダのパスの指定において、「\」を二重にしなくてはいけないという点です。

# 通常は、「C:\」で済みますが、「C:\\」という風に「\」を「\\」と表記しなくてはいけません。

# メニューバーの中から、「ファイル」→「ディレクトリの変更」としても、同様の操作を行うことができます。

ct2 <- read.csv(file="cattaneo2.csv")

# CSV形式のデータセットを読み込むためには、read.csvという関数を使います。

# fileという引数で、ファイル名を指定します。カレントのディレクトリにファイル名が一致するCSVファイルがあれば、読み込まれます。

head(ct2)

# データセットの中身を表示します。

set.seed(123456)

# 擬似乱数のシードを設定します。マッチングの結果を、事後的に正確に再現するためには、シードを設定しておく必要があります。

#------------------------------------------------------------------

# 2. 必要なパッケージの読み込み

#------------------------------------------------------------------

install.packages("rms")

install.packages("Matching")

install.packages("stddiff")

install.packages("metafor")

# 既に、インストールされている場合は、飛ばしていただいて結構です。

library(rms)

library(Matching)

library(survival)

library(lme4)

library(stddiff)

library(metafor)

#------------------------------------------------------------------

# 3. 説明変数の処理

#------------------------------------------------------------------

ct2$mmarried <- as.numeric(ct2$mmarried=="married")

ct2$fbaby <- as.numeric(ct2$fbaby=="Yes")

# 文字列で入力されているカテゴリカル変数を、数値データ（0 or 1）に変換。

ct2$mrace <- factor(ct2$mrace)

ct2$mhisp <- factor(ct2$mhisp)

ct2$foreign <- factor(ct2$foreign)

ct2$alcohol <- factor(ct2$alcohol)

ct2$mmarried <- factor(ct2$mmarried)

ct2$fbaby <- factor(ct2$fbaby)

# factorは、数値データや文字列のデータを、カテゴリ属性を持つカテゴリカル変数（factor変数）に変換する関数です。

# 以後の傾向スコアマッチングの関数にかけるためには、カテゴリカル変数は、R上でfactor変数として定義する必要があります。

#------------------------------------------------------------------

# 4. 単変量解析

#------------------------------------------------------------------

table(ct2$Y, ct2$mismoke) # 低出生時体重と母親の喫煙に関しての分割表

binom.test(95, 864) # 喫煙者群のイベント発生率

binom.test(184, 3778) # 非喫煙者群のイベント発生率

###

install.packages("fmsb")

library("fmsb")

riskdifference(95,184,864,3778)

# リスク差を計算する関数です。引数は、(1) 群１のイベント数、(2) 群２のイベント数、(3) 群１のサンプルサイズ、(4) 群２のサンプルサイズ、となります。

riskratio(95,184,864,3778)

# リスク比を計算する関数です。引数は、(1) 群１のイベント数、(2) 群２のイベント数、(3) 群１のサンプルサイズ、(4) 群２のサンプルサイズ、となります。

oddsratio(95,184,769,3778)

# オッズ比を計算する関数です。引数は、(1) 群１のイベント数、(2) 群２のイベント数、(3) 群１のイベントを起こさなかった人の人数、(4) 群２のイベントを起こさなかった人の人数、となります。

#------------------------------------------------------------------

# 5. マッチング前の背景要因の分布（サマリー表の算出）

#------------------------------------------------------------------

prs <- summary(mismoke ~ mage + medu + mmarried + mrace + mhisp + foreign + alcohol + fbaby + fage + fedu, method="reverse", overall=TRUE, test=TRUE, data=ct2)

print(prs)

# 背景要因のサマリー表の算出を行う関数です。

stddiff.numeric(ct2,gcol=16,vcol=8:11)

stddiff.binary(ct2,gcol=16,vcol=c(2,18,3,5,6,23))

# Standardized differenceの計算を行う関数です。

# 連続変数と2値変数で、別々の関数を使います。

# gcolで、治療・曝露についての変数が格納されている変数の列番号を指定します（このケースでは、ct2の16列目に喫煙の変数 mismokeが格納されています）。

# vcolで、それぞれ、Standardized differenceの計算を行いたい変数の列番号（複数でも可；ベクトルとして指定）を指定します。

#------------------------------------------------------------------

# 6. 傾向スコアの作成

#------------------------------------------------------------------

fm1 <- glm(mismoke ~ mage + medu + mmarried + mrace + mhisp + foreign + alcohol + fbaby + fage + fedu, family=binomial, data=ct2)

summary(fm1)

# 「曝露（治療）を受ける確率」を、ロジスティック回帰で推定します。

ps1 <- as.numeric(fm1$fitted)

# 「曝露（治療）を受ける確率の推定値」が、傾向スコアです。

#------------------------------------------------------------------

# 7. 傾向スコアによるマッチング

#------------------------------------------------------------------

ct2$linear.ps1 <- log(ps1/(1-ps1))

# 傾向スコアをロジット変換したもの（ロジスティック回帰モデルの線形スコア；すべての対象者において、傾向スコアと大小関係は同じです）を、マッチングの基準には使います。

caliper.opt1 <- 0.2\*sd(ct2$linear.ps1)

# キャリパーは、上記の設定が、最も一般的です。

rr1 <- Match(Y = NULL, Tr = (ct2$mismoke==1), X = ct2$linear.ps1, M=1, caliper =caliper.opt1, ties=FALSE, replace = FALSE)

summary(rr1)

# 1:1マッチングのためのコマンドです。

# 傾向スコアによるマッチングです。キャリパーの範囲内に、線形スコアが含まれる対象者同士を、1:1でマッチングします。

rr2 <- Match(Y = NULL, Tr = (ct2$mismoke==1), X = ct2$linear.ps1, M=2, caliper =caliper.opt1, ties=FALSE, replace = FALSE)

summary(rr2) # 1:2 matching

# 1:2マッチングのためのコマンドです。

# 傾向スコアによるマッチングです。キャリパーの範囲内に、線形スコアが含まれる対象者同士を、1:2でマッチングします。

###

# 以降は、単純さのために、1:2マッチングを行った場合の解析事例をお示しします（rr2をrr1に変えれば、1:1マッチングのデータの解析コマンドとなります）。

pid1 <- unique(rr2$index.treated)

pid2 <- rr2$index.control

mid1 <- 1:length(pid1)

mid2 <- sort(rep(mid1,times=2))

pid <- c(pid1,pid2)

mid <- c(mid1,mid2)

ct2\_m <- ct2[pid,]

ct2\_m$mid <- mid

# 1:2マッチングによる、Macthed datasetの作成を行うためのコマンドです。

#------------------------------------------------------------------

# 8. マッチング後の背景要因の分布（サマリー表の算出）

#------------------------------------------------------------------

pprs <- summary(mismoke ~ mage + medu + mmarried + mrace + mhisp + foreign + alcohol + fbaby + fage + fedu, method="reverse", overall=TRUE, test=TRUE, data=ct2\_m)

print(pprs)

# 背景要因のサマリー表の算出を行う関数です。

stddiff.numeric(ct2\_m,gcol=16,vcol=8:11)

stddiff.binary(ct2\_m,gcol=16,vcol=c(2,18,3,5,6,23))

# Standardized differenceの計算を行う関数です。

# 連続変数と2値変数で、別々の関数を使います。

# gcolで、治療・曝露についての変数が格納されている変数の列番号を指定します（このケースでは、ct2の16列目に喫煙の変数 mismokeが格納されています）。

# vcolで、それぞれ、Standardized differenceの計算を行いたい変数の列番号（複数でも可；ベクトルとして指定）を指定します。

#------------------------------------------------------------------

# 9. マッチング後の解析（リスク差・リスク比・オッズ比の推定）

#------------------------------------------------------------------

D <- table(ct2\_m$mid, ct2\_m$Y ,ct2\_m$mismoke)

# マッチング後のデータの集計を行ったデータセット。

a <- D[,2,2] # マッチドセットごとの喫煙者群のケースの人数。

b <- D[,1,2] # マッチドセットごとの喫煙者群の非ケースの人数。

c <- D[,2,1] # マッチドセットごとの非喫煙者群のケースの人数。

d <- D[,1,1] # マッチドセットごとの非喫煙者群の非ケースの人数。

rma.mh(ai=a, bi=b, ci=c, di=d, measure="RD")

# Mantel-Haenszel法による、マッチング後のリスク差の推定値の計算を行います。

rma.mh(ai=a, bi=b, ci=c, di=d, measure="RR")

# Mantel-Haenszel法による、マッチング後のリスク比の推定値の計算を行います。

cl1 <- clogit(Y ~ mismoke + strata(mid),data=ct2\_m)

summary(cl1)

# 条件付きロジスティック回帰による、マッチング後のオッズ比の推定値の計算を行います。