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

# 昭和大学　第10回実践臨床統計学セミナー　Rハンズオン

# 多重代入法による欠測データの解析

# 2022年3月14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。

tgc <- read.csv(file="ovariancancer.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(tgc)

# データセットの中身を表示します。

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

# 2. 必要なパッケージの読み込み

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

install.packages("mice")

# 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(mice)

library(survival)

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

# 3. 説明変数の処理

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

N <- dim(tgc)[1]

attach(tgc)

log\_ca125 <- log(ca125)

log\_alp <- log(alp)

grade <- factor(grade, levels=c("well differentiated", "moderately differentiated", "poorly differentiated"))

histol <- factor(histol, levels=c("Serous papillary", "Adenocarcinoma", "Endometrioid", "Mesonephroid (clear cell)", "Mixed mesodermal", "Mucinous", "Undifferentiated" ))

resdis <- factor(resdis, levels=c(">5 cm", "2-5 cm", "<2 cm"))

t <- tgc[,20] - tgc[,21]

d <- tgc[,19]

tgce <- data.frame(patno, age, figo, grade, histol, ascites, ps, resdis, log\_ca125, log\_alp, d, t)

detach(tgc)

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

# 4. 完全データ解析

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

ph1 <- coxph(Surv(t, d) ~ age + figo + grade + histol + ascites + ps + resdis + log\_ca125 + log\_alp, data=tgce)

summary(ph1)

# Rのデフォルトの設定では、一般的なglmやcoxphの解析では、完全データ解析が行われます。

# 欠測が1つの変数にでも含まれる対象者は解析から除外され、解析に使用できる対象者が大幅に減ります。

# 加えて、欠測メカニズムがMCARでない限り、解析の結果には、バイアスが含まれます。

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

# 5. 多重代入法による解析

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

predmt1 <- (1 - diag(1, ncol(tgce)))

predmt1[c(1,11,12),] <- predmt1[,c(1,11,12)] <- 0

# 予測モデルに含める変数の組を指定しています（1であれば含める、0であれば含めない。行列形式で指定をします）。

# 難しいようであれば、miceにかけるデータセットを、解析に含める変数のみで構成すればOKです。

imp.tgce <- mice(tgce, m=200, predictorMatrix=predmt1, seed=34871)

# m は、代入値の組の数（デフォルトでは5であるが、ここでは200としている）

# predictorMatrixで、補完モデルに含める変数の組を指定している

# methodで、代入値の生成方法を指定できる。特に指定しなければ、デフォルトの方法を採用。

# seedは、擬似乱数のシードです。多重代入法の結果を、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。

print(imp.tgce)

complete(imp.tgce, 5) # 5番目のimputed datasetを出力

ph2 <- with(imp.tgce, coxph(Surv(t, d) ~ age + figo + grade + histol + ascites + ps + resdis + log\_ca125 + log\_alp))

# 補完後のデータセット（M組）の解析；Cox回帰

pool2 <- pool(ph2)

summary(pool2)

# M組の解析結果の統合と結果を表?します。

pool3 <- summary(pool2)

HR <- exp(pool3$estimate)

CI1 <- exp(pool3$estimate - 1.96\*pool3$std.error)

CI2 <- exp(pool3$estimate + 1.96\*pool3$std.error)

P <- pool3$p.value

data.frame(pool3$term,HR,CI1,CI2,P)

# ハザード比のスケールでの解析結果