Reproducible Simulation of Boulesteix et al. (2020)

Quinty Boer

This document contains the steps and R code from a statistical simulation example given by Boulesteix et al. (2020), in a way that should be reproducible. The simulation example investigates the impact of measurement error on a continuous dependent variable (HbA1c) and confounding variable (BMI), using NHANES data. The paper and the supplemental material with the R code ccan be accessed through https://bmjopen.bmj.com/content/10/12/e039921. Except for minor changes to file paths and the code to saving the plot (using `ggsave` to PNG instead of `savePlot` to TIFF), the original code is unaltered.

Load libraries and read in data

The first steps include loading the required libraries `Hmisc`, `mice`, and `tidyverse`, as well as reading in the data. Details on the data can be found at https://wwwn.cdc.gov/nchs/nhanes/, and the complete database can be assessed through https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Begi-nYear=2015. For this simulation, the data files were downloaded directly from https://github.com/gerkovink/markup/blob/main/documents/week-4/raw_data.

```
# load libraries
library(Hmisc)
library(mice)
library(tidyverse)
```

```
# read data
d1 <- sasxport.get("../data/DEMO_I.xpt")
d2 <- sasxport.get("../data/BPX_I.xpt")
d3 <- sasxport.get("../data/BMX_I.xpt")
d4 <- sasxport.get("../data/GHB_I.xpt")
d5 <- sasxport.get("../data/TCHOL_I.xpt")
```

```
d1.t <- subset(d1,select=c("seqn","riagendr","ridageyr"))
d2.t <- subset(d2,select=c("seqn","bpxsy1"))
d3.t <- subset(d3,select=c("seqn","bmxbmi"))
d4.t <- subset(d4,select=c("seqn","lbxgh"))
d5.t <- subset(d5,select=c("seqn","lbdtcsi"))
d <- merge(d1.t,d2.t)
d <- merge(d,d3.t)
d <- merge(d,d4.t)
d <- merge(d,d5.t)
```

```
# rename variables:
# RIAGENDR - Gender
# RIDAGEYR - Age in years at screening
# BPXSY1 - Systolic: Blood pres (1st rdg) mm Hg
# BMXBMI - Body Mass Index (kg/m**2)
# LBDTCSI - Total Cholesterol (mmol/L)
# LBXGH - Glycohemoglobin (%)
d$age <- d$ridageyr
d$sex <- d$riagendr
d$bp <- d$bpxsy1
d$bmi <- d$bmxbmi
d$HbA1C <- d$lbxgh
d$chol <- d$lbdtcsi
d$age[d$age<18] <- NA
```

```
# select complete cases:
dc <- cc(subset(d,select=c("age","sex","bmi","HbA1C","bp")))
# analysis:
summary(lm(bp ~ HbA1C + age + as.factor(sex), data=dc))
confint(lm(bp ~ HbA1C + age + as.factor(sex), data=dc))
summary(lm(bp ~ HbA1C + bmi + age + as.factor(sex), data=dc))
confint(lm(bp ~ HbA1C + bmi + age + as.factor(sex), data=dc))
```

Run simulations

In the simulation example from Boulesteix et al. (2020), the original observations are assumed to be measured without error. New variables are created through simulation where measurement error is added to the variables HbA1c and/or BMI, with errors drawn from a normal distribution. The amount of measurement errors differs in the simulation scenarios. Each scenario is repeated 1000 times and the results are then averaged (Boulesteix et al. 2020).

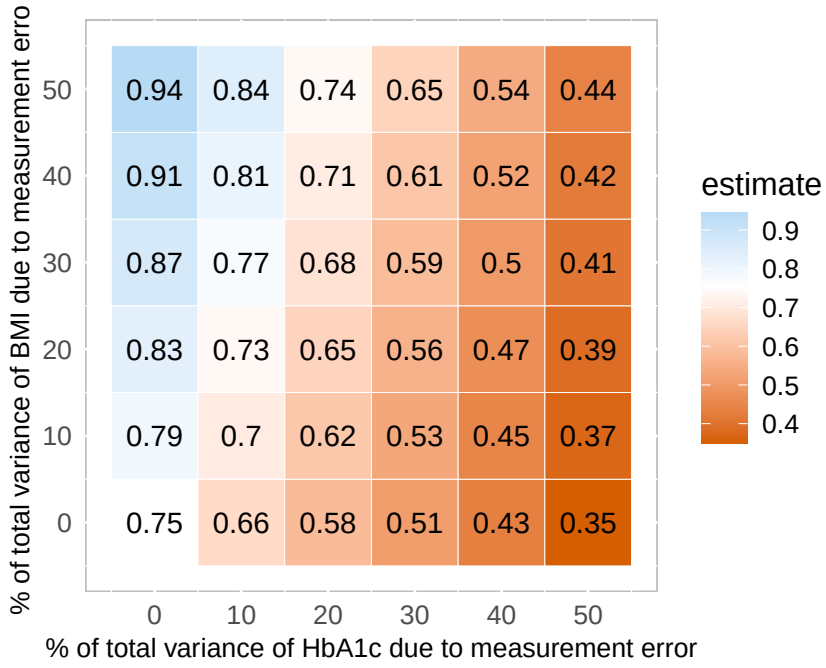
```r
# simulation of measurement error:
ref <- lm(bp ~ HbA1C + bmi + age + as.factor(sex), data=dc)$coef[2]
n.sim <- 1e3
perc.me.exp <- seq(0,.5,.1)
perc.me.conf<- seq(0,.5,.1)
scenarios <- expand.grid(perc.me.exp,perc.me.conf)
var.exp <- var(dc$HbA1C)
var.conf <- var(dc$bmi)
n <- dim(dc)[1]
beta.hat <- matrix(ncol=dim(scenarios)[1], nrow=n.sim)
for (k in 1:n.sim){
  #print(k)
  set.seed(k)
  for (i in 1:dim(scenarios)[1]){
    var.me.exp <- var.exp*scenarios[i,1]/(1-scenarios[i,1])
    var.me.conf <- var.conf*scenarios[i,2]/(1-scenarios[i,2])
    dc$HbA1C.me <- dc$HbA1C + rnorm(dim(dc)[1], 0, sqrt(var.me.exp) )
    dc$bmi.me <- dc$bmi + rnorm(dim(dc)[1], 0, sqrt(var.me.conf) )
    beta.hat[k,i] <- lm(bp ~ HbA1C.me + age + bmi.me + as.factor(sex), data=dc)$coef[2]
  }}
```

Produce plot

The code below produces the Figure 2 in the paper by @Boulesteix et al. (2020). It shows the estimates of the correlation between HbA1c and blood pressure when controlling for BMI in the different simulation scenarios.

```r
# create figure:
tot.mat <- cbind(100*scenarios,apply(beta.hat,2,mean))
colnames(tot.mat) <- c("me.exp","me.conf","estimate")
FIGURE <- ggplot(tot.mat, aes(me.exp, me.conf)) +
  geom_tile(color="white",aes(fill = estimate)) +
  geom_text(aes(label = round(estimate, 2))) +
  scale_fill_gradient2(low="#D55E00",mid="white",high = "#56B4E9", midpoint=ref) +
  labs(x=paste("% of total variance of HbA1c due to measurement error"),
       y=paste("% of total variance of BMI due to measurement error")) +
  coord_equal()+
  scale_y_continuous(breaks=unique(tot.mat[,1]))+
  scale_x_continuous(breaks=unique(tot.mat[,1]))+
  theme(panel.background = element_rect(fill='white', colour='grey'),
        plot.title=element_text(hjust=0),
```

```
        axis.ticks=element_blank(),
        axis.title=element_text(size=10),
        axis.text=element_text(size=10),
        legend.title=element_text(size=12),
        legend.text=element_text(size=10))
FIGURE
```




```
# savePlot("../results/Figure_STRATOS.tif", type="tif")
ggsave(filename = "../results/Figure_STRATOS.png", plot = FIGURE)
```

References

Boulesteix, Anne-Laure, Rolf HH Groenwold, Michal Abrahamowicz, Harald Binder, Matthias Briel, Roman Hornung, Tim P Morris, Jörg Rahnenführer, and Willi Sauerbrei. 2020. "Introduction to Statistical Simulations in Health Research." *BMJ Open* 10 (12): e039921. https://doi.org/10.1136/bmjopen-2020-039921.