My Reproducible Manuscript

Adam Maghout

Purpose of the project

The project presented here is a reproduction of work by Boulesteix et al. (2020), done for the purpose of an exercise of the Markup Languages and Reproducible Programming in Statistics course offered by Utrecht University. The code used here was provided by the original researchers in pdf form then converted into workable R code by Gerko Vink and Hanne Oberman. Likewise, the data used was also extracted from the website of the Centers for Disease Control and Prevention (CDC) by these two researchers. Presenting the findings and formatting this document were done by Adam Maghout.

Libraries

The required libraries for this project are given below:

```
library(Hmisc)
library(mice)
library(tidyverse)
```

Data

Data from 5092 subjects in the 2015–2016 National Health and Nutrition Examination Survey (NHANES) are used to obtain an estimate of the effect of HbA1c on systolic blood pressure, while adjusting for age, gender and body mass index (BMI). Since the NHANES website provides the data in several datasets, we must first select only the variables of interest for our analysis. These are:

- 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 (%)

Information on other variables in the datasets is given on the NHANES website but also in the column descriptions in the datasets.

```
# Read data:
d1 <- sasxport.get("../data/DEMO_I.xpt")
```

```
Processing SAS dataset DEMO_I    ..
```

```
d2 <- sasxport.get("../data/BPX_I.xpt")
```

```
Processing SAS dataset BPX_I     ..
```

```
d3 <- sasxport.get("../data/BMX_I.xpt")
```

```
Processing SAS dataset BMX_I     ..
```

```
d4 <- sasxport.get("../data/GHB_I.xpt")
```

```
Processing SAS dataset GHB_I     ..
```

```
d5 <- sasxport.get("../data/TCHOL_I.xpt")
```

```
Processing SAS dataset TCHOL_I   ..
```

```
# Filter for variables of interest
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)
```

```r
d <- merge(d,d5.t)

# Rename variables:
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")))
```

Baseline Model

The effect of HbA1c on systolic blood pressure is measured using a linear model. Initially, this model is given whilst adjusting for age and sex. Then, BMI is also used as an adjusting factor:

```r
# Model 1
summary(lm(bp ~ HbA1C + age + as.factor(sex), data=dc))
```

```
Call:
lm(formula = bp ~ HbA1C + age + as.factor(sex), data = dc)

Residuals:
Systolic: Blood pres (1st rdg) mm Hg
    Min      1Q  Median      3Q     Max
-49.887 -10.509  -1.378   8.491 107.583

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     98.75149    1.21418  81.332  < 2e-16 ***
HbA1C            1.12638    0.20291   5.551 2.98e-08 ***
age              0.44486    0.01284  34.648  < 2e-16 ***
as.factor(sex)2 -3.24792    0.45164  -7.191 7.34e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 16.1 on 5088 degrees of freedom
Multiple R-squared:  0.2305,    Adjusted R-squared:   0.23
F-statistic:   508 on 3 and 5088 DF,  p-value: < 2.2e-16
```

```r
confint(lm(bp ~ HbA1C + age + as.factor(sex), data=dc))
```

```
                     2.5 %      97.5 %
(Intercept)     96.3711755 101.1317982
HbA1C            0.7285836   1.5241825
age              0.4196932   0.4700355
as.factor(sex)2 -4.1333281  -2.3625106
```

In this first model, we find that HbA1c increases systolic blood pressure by 1.13 mmHg (95% CI 0.73 to 1.52) per unit increase in HbA1c. This is given under the coefficients in the summary.

```r
# Model 2
summary(lm(bp ~ HbA1C + bmi + age + as.factor(sex), data=dc))
```

```
Call:
lm(formula = bp ~ HbA1C + bmi + age + as.factor(sex), data = dc)

Residuals:
Systolic: Blood pres (1st rdg) mm Hg
    Min      1Q  Median      3Q     Max
-51.068 -10.251  -1.504   8.264 107.410

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     92.65583    1.39320  66.506  < 2e-16 ***
HbA1C            0.75177    0.20596   3.650 0.000265 ***
bmi              0.28632    0.03282   8.724  < 2e-16 ***
age              0.44586    0.01275  34.979  < 2e-16 ***
as.factor(sex)2 -3.63115    0.45049  -8.060  9.4e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.98 on 5087 degrees of freedom
Multiple R-squared:  0.2418,    Adjusted R-squared:  0.2412
F-statistic: 405.7 on 4 and 5087 DF,  p-value: < 2.2e-16
```

```r
confint(lm(bp ~ HbA1C + bmi + age + as.factor(sex), data=dc))
```

```
                     2.5 %     97.5 %
(Intercept)     89.9245592 95.3871089
HbA1C            0.3479966  1.1555348
bmi              0.2219815  0.3506673
age              0.4208695  0.4708464
as.factor(sex)2 -4.5143014 -2.7479929
```

In this second model, adjusting for BMI, we find that HbA1c instead increases systolic blood pressure by 0.75 mmHg (95% CI 0.35 to 1.16) per unit increase in HbA1c.

Added Measurement error

Measurement error is added to the HbA1c and BMI data by setting a proportion of the observed variance in both variables to be due to measurement error. This proportion ranges from 0 to 50%, yielding a total of $6 * 6 = 30$ scenarios across both variables. This process is repeated 1 000 times for each scenario to obtain measures of spread.

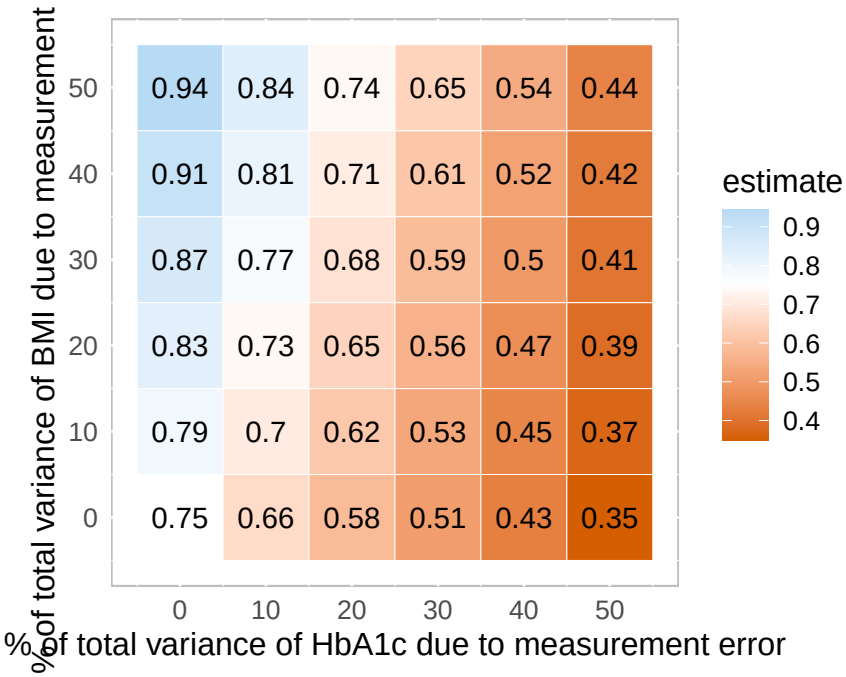
```r
# simulation of measurement error:
# Baseline increase
ref <- lm(bp ~ HbA1C + bmi + age + as.factor(sex), data=dc)$coef[2]

n.sim <- 1e3 # Number of iterations

# exp = exposure variable; conf = confounding variable
perc.me.exp <- seq(0,.5,.1) # Proportion of measurement error variance for the HbA1c
perc.me.conf<- seq(0,.5,.1) # Proportion of measurement error variance for the BMI
scenarios <- expand.grid(perc.me.exp,perc.me.conf) # 36 scenarios

# Variances without measurement error
var.exp <- var(dc$HbA1C)
var.conf <- var(dc$bmi)

n <- dim(dc)[1] # Number of observations

beta.hat <- matrix(ncol=dim(scenarios)[1], nrow=n.sim) # Initialise results matrix

# Simulate measurement error
for (k in 1:n.sim){
```

```
  # print(k) # Can be used for tracking
  set.seed(k)
  for (i in 1:dim(scenarios)[1]){ # 1 000 iterations
    # Compute variances
    var.me.exp <- var.exp*scenarios[i,1]/(1-scenarios[i,1])
    var.me.conf <- var.conf*scenarios[i,2]/(1-scenarios[i,2])

    # Add meaurement error
    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) )

    # Obtain desired coefficient
    beta.hat[k,i] <- lm(bp ~ HbA1C.me + age + bmi.me + as.factor(sex), data=dc)$coef[2]
  }}
```

Figure

Given the results from the previous section, figure 2 is obtained as follows:

```
# 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=12),
        axis.text=element_text(size=10),
        legend.title=element_text(size=12),
        legend.text=element_text(size=10))

FIGURE # display the figure
```




```
# Save the figure (optional)
# savePlot("Figure_STRATOS.tif", type="tif")
```

The relation between HbA1c and systolic blood pressure was attenuated when measurement error was added to HbA1c, but not when measurement error was added to BMI. The association became stronger as measurement error was added solely to the confounding variable BMI. The reason for this effect is that, with increasing levels of measurement error on BMI, adjustment for the confounding due to BMI becomes less efficient and the effect estimate gets closer to the unadjusted estimate (1.13 mmHg). Due to measurement error, a type of residual confounding is introduced. In the case of measurement error on HbA1c as well as BMI, both phenomena play a role and may cancel each other out.

Bibliography

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.