Supplementary Digital Content to 'Eliminating survivor bias in two-stage instrumental variable estimators'

Stijn Vansteelandt[a,b], Stefan Walter[c,d] and Eric Tchetgen Tchetgen[e]

From the [a]Department of Applied Mathematics, Computer Sciences and Statistics, Ghent University, Ghent, Belgium; [b]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK; [c]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA; [d]Fundación de Investigación Biomédica, Hospital Universitario de Getafe, Madrid, Spain; and [e]Deptartment of Statistics, The Wharton School, University of Pennsylvania, PA.

eAppendix A: asymptotic unbiasedness of the two-stage estimator

Let $R_i(t) \equiv I(T_i \geq t)$, $dN_i(t) = I(T_i = t)$ and $R_{i0}(t) \equiv I(T_{i0} \leq t)$. Following a similar reasoning as in[1], one can see that the suggested two-stage estimator is asymptotically equivalent to the solution to the following estimating equation:

$$0 = \sum_{i=1}^{n} \int_0^{\infty} \left[M_i(t) - E\left\{M_i(t) | T_i \geq t, T_{i0} < t\right\}\right] R_i(t) R_{i0}(t) \left\{dN_i(t) - \psi M_i(t) dt - d\Omega_0(t)\right\}.$$

When $M_i(t)$ is obtained by fitting model

$$E(A_i | Z_i, T_i \geq t, T_{i0} < t) = \alpha_0(t) + \alpha_1(t) Z_i,$$

using ordinary least squares, then $\alpha_0(t)$ and $\alpha_1(t)$ are obtained as solutions to

$$0 = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ Z_i \end{pmatrix} R_i(t) R_{i0}(t) \left\{ A_i - M_i(t) \right\}. \tag{1}$$

Under the considered additive hazard model, Equation 3 of the main paper, the contributions to the estimating equation for ψ then have mean

$$E\left(\int_0^\infty \left[M_i(t) - E\left\{ M_i(t) | T_i \geq t, T_{i0} < t \right\} \right] R_i(t) R_{i0}(t) \left[\psi \left\{ A_i - M_i(t) \right\} dt + d\Omega(t, U_i) \right] \right),$$

for some function $\Omega(t, U_i)$, which reduces to

$$E\left(\int_0^\infty \left[M_i(t) - E\left\{ M_i(t) | T_i \geq t, T_{i0} < t \right\} \right] R_i(t) R_{i0}(t) d\Omega(t, U_i) \right),$$

by the ordinary least squares restrictions (1), regardless of whether the first-stage model is correctly specified. The latter expectation equals zero under the restrictions in Equation 5 of the main paper. Note that these restrictions are satisfied under the additive hazards model, Equation 3 of the main paper, when

$$A_i - E(A_i | Z_i) \perp\!\!\!\perp U_i$$

and $T_{i0} \perp\!\!\!\perp (Z_i, A_i, T_i) | U_i$ (see[2]). The latter assumption is guaranteed to hold when there is no birth cohort effect on the genotype distribution, in the sense that Z is independent of T_0 (i.e., when the population allele frequencies are the same across birth cohorts), for then one can let U be a vector of variables that includes T_0 (see[2]).

eAppendix B: R code

The proposed analysis can be conducted based on the following R code, where `ti` is the vector of observed event (or censoring) times, `t0` is the vector of observed entry times, `d` the event indicator (1 if an event occurred, 0 if censoring occurred), `a` the exposure, `z` the IV and `id` a subject identifier. We provide an artificial, simulated dataset to enable the user to perform a test run of the code.

```
install.packages("timereg")
library(timereg)

# generate artificial test dataset

n <- 500
u <- rnorm(n)
z <- rbinom(n,1,0.5)
a <- rnorm(n,3+z-u)
ti <- rexp(n,1/abs(0.5*a+0.25*u))
d <- ifelse(ti<10,1,0)
ti <- pmin(ti,10)
t0 <- runif(n,0,0.5)
id <- 1:length(a)

# vector of observed event times

times <- sort(unique(c(t0,ti)[t0<ti]))
n <- length(times)

# construct longitudinal dataset which expresses all observed at risk periods

dataset <- data.frame(y=rep(y,each=n),ti=rep(ti,each=n),t0=rep(t0,each=n),
   d=rep(d,each=n),a=rep(a,each=n),z=rep(z,each=n),id=rep(id,each=n),
   start=rep(c(0,times[-n]),length=n),stop=rep(times,length=n))
dataset <- dataset[dataset$t0<dataset$ti,]
dataset <- dataset[dataset$t0<=dataset$start,]
dataset <- dataset[dataset$ti>=dataset$stop,]
dataset$d[dataset$ti>dataset$stop] <- 0

# construct predictions M(t) - denoted m

dataset$m <- NULL
for (i in 1:(length(times)-1)){
ty <- times[i]
s <- (dataset$ti >= ty)&(dataset$t0 <= ty)
dataset$m[dataset$start==ty] <- predict(lm(a~z, data=dataset[s,]),
   newdata=data.frame(z=dataset$z[dataset$start==ty]))
}

# fit the second-stage additive hazards model

aalen(Surv(start,stop,d)~const(m),data = dataset)
```

We moreover provide code for assessing the plausibility of the location shift assumption.

```
res <- matrix(0,ncol=3,nrow=100)
delta <- resid(lm(a~z))
for (i in 1:100){
# choose a range of values theta up to the maximum accumulated exposure effect
# here an exposure effect of 0.08 which accumulates for at most 30 years
res[i,1] <- (30*i/100)*0.08
# calculate the correlation
res[i,2] <- cor(z,exp(-res[i,1]*delta))
# test for evidence of a correlation
res[i,3] <- cor.test(z,exp(-res[i,1]*delta))$p.value
}
par(mfrow=c(1,2))
plot(res[,1],res[,2],xlab=expression(theta),ylab="Correlation")
plot(res[,1],log(res[,3]),xlab=expression(theta),ylab="log p-value",
   ylim=c(log(0.001),0))
abline(h=log(0.05))
```

Significant evidence of a correlation is suggestive of a violation of the location shift assumption.

References

1. Vansteelandt S, Martinussen T, Tchetgen E. On adjustment for auxiliary covariates in additive hazard models for the analysis of randomized experiments. Biometrika. 2014;101(1).

2. Vansteelandt S, Dukes O, Martinussen T. Survivor bias in Mendelian randomisation analysis. Biostatistics. 2017;https://doi.org/10.1093/biostatistics/kxx050.