It has been written about *within-* and *between-period* intra-cluster correlations in the context of stepped-wedge cluster randomized study designs. These are quite important to understand when figuring out sample size requirements (and models for analysis, which I’ll be writing about soon.) Here, We are extending the constant ICC assumption that’s presented last time around by introducing some complexity into the correlation structure.

ICC(Intra Cluster Correlation)

**Intra-cluster correlations**

If outcomes of subjects in a study are correlated in any way (e.g. they received care from the same health care provider), we do not learn as much information from each individual study participant as we would in the case where there is no correlation. In a parallel designed cluster randomized trial (where half of the clusters receive an intervention and the other half do not), we expect that the outcomes will be correlated *within* each cluster, though not *across* clusters. (This is not true if the clusters are themselves clustered, in which case we would have a 2-level clustered study.) This intra-cluster correlation (ICC) increases sample size requirements and reduces precision/power.

A common way to model correlation explicitly in a cluster randomized trial is to conceive of a random effects model like this:

(1) \qquad \qquad Y\_{ic} = \mu + \beta\_1X\_{c} + b\_c + e\_{ic},(1)*Yic*​=*μ*+*β*1​*Xc*​+*bc*​+*eic*​,

where Y\_{ic}*Yic*​ is a continuous outcome for subject i*i* in cluster c*c*, and X\_c*Xc*​ is a treatment indicator for cluster c*c* (either 0 or 1). The underlying structural parameters are \mu*μ*, the grand mean, and \beta\_1*β*1​, the treatment effect. The unobserved random effects are, b\_c \sim N(0, \sigma^2\_b)*bc*​∼*N*(0,*σb*2​), the normally distributed group level effect, and e\_{ic} \sim N(0, \sigma^2\_e)*eic*​∼*N*(0,*σe*2​), the normally distributed individual-level effect. (This is often referred to as the “error” term, but that doesn’t adequately describe what is really unmeasured individual variation.)

The correlation between any two subjects i*i* and j*j* in the *same* cluster c*c* is:

cor(Y\_{ic}, Y\_{jc}) = \frac{cov(Y\_{ic}, Y\_{jc})} {\sqrt {var(Y\_{ic})var(Y\_{jc})}}*cor*(*Yic*​,*Yjc*​)=*var*(*Yic*​)*var*(*Yjc*​)​*cov*(*Yic*​,*Yjc*​)​

cov(Y\_{ic}, Y\_{jc})*cov*(*Yic*​,*Yjc*​) can be written in terms of the parameters in the underlying data generating process:

\begin{aligned} cov(Y\_{ic}, Y\_{jc}) &amp;= cov(\mu + \beta\_1X\_c + b\_c + e\_{ic}, \mu + \beta\_1X\_c + b\_c + e\_{jc}) \\ &amp;=cov(b\_c, b\_c) + cov(e\_{ic},e\_{jc} ) \\ &amp;=\sigma^2\_b + 0 \\ &amp;=\sigma^2\_b \end{aligned}*cov*(*Yic*​,*Yjc*​)​=*cov*(*μ*+*β*1​*Xc*​+*bc*​+*eic*​,*μ*+*β*1​*Xc*​+*bc*​+*ejc*​)=*cov*(*bc*​,*bc*​)+*cov*(*eic*​,*ejc*​)=*σb*2​+0=*σb*2​​

The terms simplify since the cluster level effects are independent of the individual level effects (and all the fixed effects in the model) and the individual level effects are independent of each other. The within-period intra-cluster co-variance depends only on the between cluster variation.

The total variance of the outcomes Y\_{ic}*Yic*​ is:

\begin{aligned} var(Y\_{ic}) &amp;= var(\mu + \beta\_1X\_c + b\_c + e\_{ic}) \\ &amp;= var(b\_c) + var(e\_{ic}) \\ &amp;= \sigma^2\_b + \sigma^2\_e \end{aligned}*var*(*Yic*​)​=*var*(*μ*+*β*1​*Xc*​+*bc*​+*eic*​)=*var*(*bc*​)+*var*(*eic*​)=*σb*2​+*σe*2​​

Substituting all of this into the original equation gives us the intra-cluster correlation for any two subjects in the cluster:

\begin{aligned} cor(Y\_{ic}, Y\_{jc}) &amp;= \frac{cov(Y\_{ic}, Y\_{jc})} {\sqrt {var(Y\_{ic})var(Y\_{jc})}} \\ \\ ICC &amp;= \frac{\sigma^2\_b}{\sigma^2\_b + \sigma^2\_e} \end{aligned}*cor*(*Yic*​,*Yjc*​)*ICC*​=*var*(*Yic*​)*var*(*Yjc*​)​*cov*(*Yic*​,*Yjc*​)​=*σb*2​+*σe*2​*σb*2​​​

So, the correlation between any two subjects in a cluster increases as the variation *between* clusters increases.

**Cluster randomization when time matters**

Moving beyond the parallel design to the stepped-wedge design, time starts to play a very important role. It is important to ensure that we do not confound treatment and time effects; we have to be careful that we do not attribute the general changes over time to the intervention. This is accomplished by introducing a time trend into the model. (Actually, it seems more common to include a time-specific effect so that each time period has its own effect. However, for simulation purposes, I will will assume a linear trend.)

In the stepped-wedge design, we are essentially estimating within-cluster treatment effects by comparing the cluster with itself pre- and post-intervention. To estimate sample size and precision (or power), it is no longer sufficient to consider a single ICC, because there are now multiple ICC’s - the within-period ICC and the between-period ICC’s. The within-period ICC is what we defined in the parallel design (since we effectively treated all observations as occurring in the same period.) Now we also need to consider the expected correlation of two individuals in the *same* cluster in *different* time periods.

If we do not properly account for within-period ICC and the between-period ICC’s in either the planning or analysis stages, we run the risk of generating biased estimates.

My primary aim is to describe possible data generating processes for the stepped wedge design and what implications they have for both the within-period and between-period ICC’s. I will generate data to confirm that observed ICC’s match up well with the theoretical expectations. This week I will consider the simplest model, one that is frequently used but whose assumptions may not be realistic in many applications. In a follow-up post, I will consider more flexible data generating processes.

**Constant ICC’s over time**

Here is probably the simplest model that can be conceived for a process underlying the stepped-wedge design:

(2) \qquad \qquad Y\_{ict} = \mu + \beta\_0t + \beta\_1X\_{ct} + b\_c + e\_{ict}(2)*Yict*​=*μ*+*β*0​*t*+*β*1​*Xct*​+*bc*​+*eict*​

As before, the unobserved random effects are b\_c \sim N(0, \sigma^2\_b)*bc*​∼*N*(0,*σb*2​) and e\_{ict} \sim N(0, \sigma^2\_e)*eict*​∼*N*(0,*σe*2​). The key differences between this model compared to the parallel design is the time trend and time-dependent treatment indicator. The time trend accounts for the fact that the outcome may change over time regardless of the intervention. And since the cluster will be in both the control and intervention states we need to have an time-dependent intervention indicator. (This model is a slight variation on the *Hussey and Hughes* model, which includes a time-specific effect \beta\_t*βt*​ rather than a linear time trend. This paper by *Kasza et al* describes this stepped-wedge model, and several others, in much greater detail.)

The *within-period* ICC from this is model is:

\begin{aligned} cor(Y\_{ict}, Y\_{jct}) &amp;= cor(\mu + \beta\_0t + \beta\_1X\_{ct} + b\_c + e\_{ict}, \ \mu + \beta\_0t + \beta\_1X\_{ct} + b\_c + e\_{jct}) \\ \\ ICC\_{tt}&amp;= \frac{\sigma^2\_b}{\sigma^2\_b + \sigma^2\_e} \end{aligned}*cor*(*Yict*​,*Yjct*​)*ICCtt*​​=*cor*(*μ*+*β*0​*t*+*β*1​*Xct*​+*bc*​+*eict*​, *μ*+*β*0​*t*+*β*1​*Xct*​+*bc*​+*ejct*​)=*σb*2​+*σe*2​*σb*2​​​

I have omitted the intermediary steps, but the logic is the same as in the parallel design case. The within-period ICC under this model is also the same as the ICC in the parallel design.

More importantly, in this case the *between-period* ICC turns out to be the same as the *within-period* ICC. For the *between-period* ICC, we are estimating the expected correlation between any two subjects i*i* and j*j* in cluster c*c*, one in time period t*t* and the other in time period t^\prime*t*′ (t \ne t^\prime)(*t*̸​=*t*′):

\begin{aligned} cor(Y\_{ict}, Y\_{jct^\prime}) &amp;= cor(\mu + \beta\_0t + \beta\_1X\_{ct} + b\_c + e\_{ict}, \ \mu + \beta\_0t^\prime + \beta\_1X\_{ct^\prime} + b\_c + e\_{jct^\prime}) \\ \\ ICC\_{tt^\prime}&amp;= \frac{\sigma^2\_b}{\sigma^2\_b + \sigma^2\_e} \end{aligned}*cor*(*Yict*​,*Yjct*′​)*ICCtt*′​​=*cor*(*μ*+*β*0​*t*+*β*1​*Xct*​+*bc*​+*eict*​, *μ*+*β*0​*t*′+*β*1​*Xct*′​+*bc*​+*ejct*′​)=*σb*2​+*σe*2​*σb*2​​​

Under this seemingly reasonable (and popular) model, we are making a big assumption that the within-period ICC and between-period ICC’s are equal and constant throughout the study. This may or may not be reasonable - but it is important to acknowledge the assumption and to make sure we justify that choice.

**Generating data to simulate a stepped-wedge design**

I’ve generated data from a stepped-wedge design before on this blog, but will repeat the details here. For the data definitions, we define the variance of the cluster-specific effects, the cluster sizes, and the outcome model.

defc <- defData(varname = "ceffect", formula = 0, variance = 0.15,

dist = "normal", id = "cluster")

defc <- defData(defc, "m", formula = 10, dist = "nonrandom")

defa <- defDataAdd(varname = "Y",

formula = "0 + 0.10 \* period + 1 \* rx + ceffect",

variance = 2, dist = "normal")

The data generation follows this sequence: cluster data, temporal data, stepped-wedge treatment assignment, and individual (within cluster) data:

dc <- genData(100, defc)

dp <- addPeriods(dc, 7, "cluster")

dp <- trtStepWedge(dp, "cluster", nWaves = 4, lenWaves = 1, startPer = 2)

dd <- genCluster(dp, cLevelVar = "timeID", "m", "id")

dd <- addColumns(defa, dd)

dd

## cluster period ceffect m timeID startTrt rx id Y

## 1: 1 0 -0.073 10 1 2 0 1 -2.12

## 2: 1 0 -0.073 10 1 2 0 2 -1.79

## 3: 1 0 -0.073 10 1 2 0 3 1.53

## 4: 1 0 -0.073 10 1 2 0 4 -1.44

## 5: 1 0 -0.073 10 1 2 0 5 2.25

## ---

## 6996: 100 6 0.414 10 700 5 1 6996 1.28

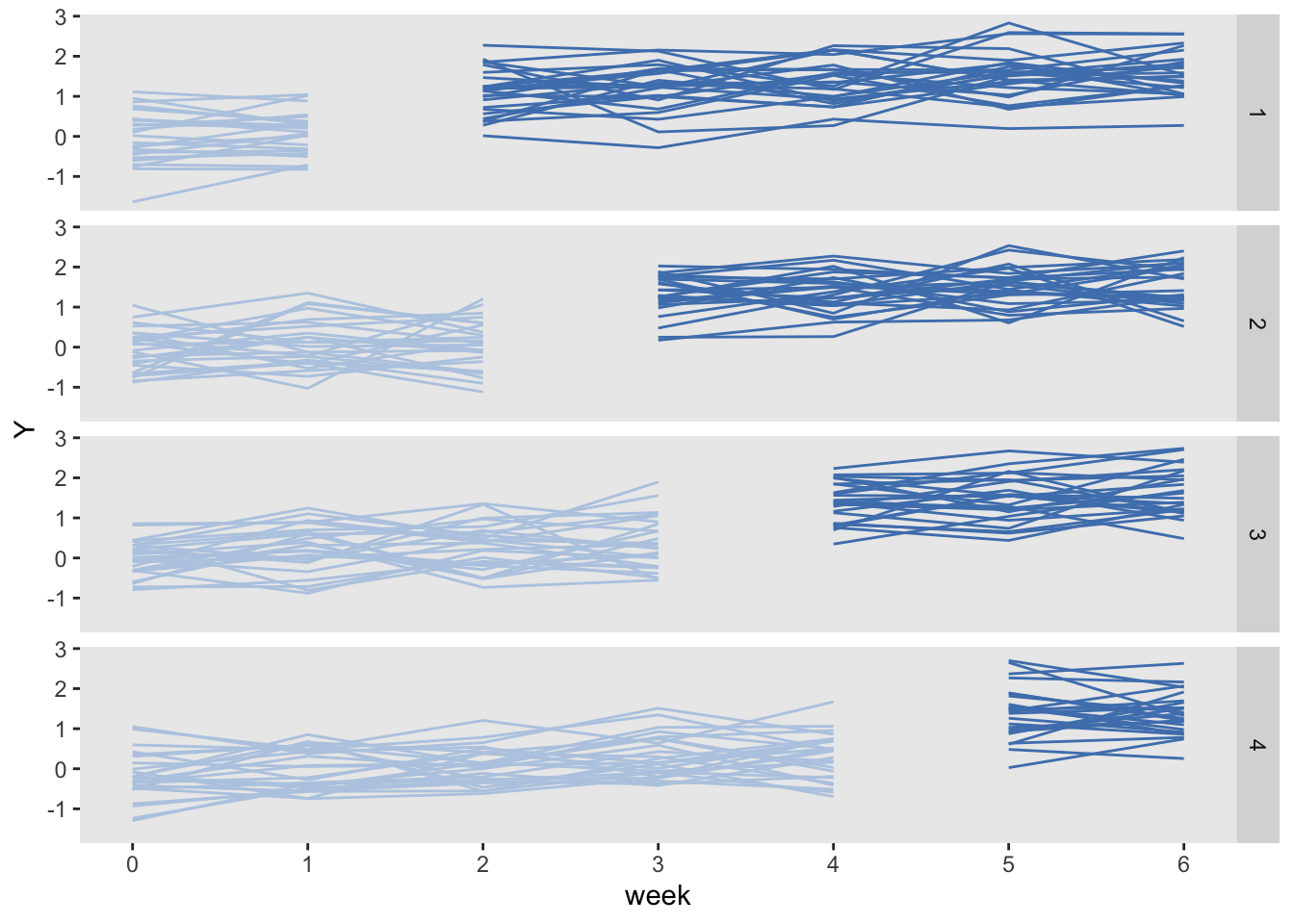
## 6997: 100 6 0.414 10 700 5 1 6997 0.30

## 6998: 100 6 0.414 10 700 5 1 6998 0.94

## 6999: 100 6 0.414 10 700 5 1 6999 1.43

## 7000: 100 6 0.414 10 700 5 1 7000 0.58

It is always useful (and important) to visualize the data (regardless of whether they are simulated or real). This is the summarized cluster-level data. The clusters are grouped together in waves defined by starting point. In this case, there are 25 clusters per wave. The light blue represents pre-intervention periods, and the dark blue represents intervention periods.



**Estimating the between-period within-cluster correlation**

I want to estimate the observed between-period within cluster correlation without imposing any pre-conceived structure. In particular, I want to see if the data generated by the process defined in equation (2) above does indeed lead to constant within- and between-period ICC’s. In a future post, I will estimate the ICC using a model, but for now, I’d prefer to estimate the ICC’s directly from the data.

A 1982 paper by *Bernard Rosner* provides a non-parametric estimate of the *between-period* ICC. He gives this set of equations to find the correlation coefficient \rho\_{tt^\prime}*ρtt*′​ for two time periods t*t* and t^\prime*t*′. In the equations, m\_{ct}*mct*​ represents the cluster size for cluster c*c* in time period t*t*, and K*K* represents the number of clusters:

\rho\_{tt^\prime} = \frac{\sum\_{c=1}^K \sum\_{i=1}^{m\_{ct}} \sum\_{j=1}^{m\_{ct^\prime}} (Y\_{ict}-\mu\_t)(Y\_{jct^\prime}-\mu\_{t^\prime})} {\left[ \left ( \sum\_{c=1}^K m\_{ct^\prime} \sum\_{i=1}^{m\_{ct}} (Y\_{ict}-\mu\_t)^2 \right ) \left ( \sum\_{c=1}^K m\_{ct} \sum\_{j=1}^{m\_{ct^\prime}} (Y\_{jct^\prime}-\mu\_{t^\prime})^2 \right )\right] ^ \frac {1}{2}}*ρtt*′​=[(∑*c*=1*K*​*mct*′​∑*i*=1*mct*​​(*Yict*​−*μt*​)2)(∑*c*=1*K*​*mct*​∑*j*=1*mct*′​​(*Yjct*′​−*μt*′​)2)]21​∑*c*=1*K*​∑*i*=1*mct*​​∑*j*=1*mct*′​​(*Yict*​−*μt*​)(*Yjct*′​−*μt*′​)​

\mu\_t = \frac{\sum\_{c=1}^K m\_{ct} m\_{ct^\prime} \mu\_{ct}}{\sum\_{c=1}^K m\_{ct} m\_{ct^\prime}} \ \ , \ \ \mu\_{t^\prime} = \frac{\sum\_{c=1}^K m\_{ct} m\_{ct^\prime} \mu\_{ct^\prime}}{\sum\_{c=1}^K m\_{ct} m\_{ct^\prime}}*μt*​=∑*c*=1*K*​*mct*​*mct*′​∑*c*=1*K*​*mct*​*mct*′​*μct*​​  ,  *μt*′​=∑*c*=1*K*​*mct*​*mct*′​∑*c*=1*K*​*mct*​*mct*′​*μct*′​​

\mu\_{ct} = \frac{\sum\_{i=1}^{m\_{ct}} Y\_{ict}}{m\_{ct}} \ \ , \ \ \mu\_{ct^\prime} = \frac{\sum\_{j=1}^{m\_{ct^\prime}} Y\_{jct^\prime}}{m\_{ct^\prime}}*μct*​=*mct*​∑*i*=1*mct*​​*Yict*​​  ,  *μct*′​=*mct*′​∑*j*=1*mct*′​​*Yjct*′​​

I’ve implemented the algorithm in R, and the code is included in the addendum. One issue that came up is that as the intervention is phased in over time, the treatment effect is present for each at different times. The algorithm breaks down as a result. However, the between-period ICC can be calculated for each wave, and then we can average across the four waves.

The *within-period* ICC is estimated using a linear mixed effects model applied to each period separately, so that we estimate period-specific within-period ICC’s. The expected (constant) ICC is 0.07 = \left(\frac{0.15}{0.15 + 2}\right)0.07=(0.15+20.15​).

The function iccs (shown below in the addendum) returns both the estimated *within-* and *between-cluster* ICC’s for a single data set. Here is the within-period ICC for the first period (actually period 0) and the between-period ICC’s using period 0:

set.seed(47463)

iccs(dd, byWave = T)[,c(22, 0:6)]

## wp0 bp01 bp02 bp03 bp04 bp05 bp06

## 1: 0.041 0.068 0.073 0.08 0.067 0.054 0.053

ICC estimates are quite variable and we can’t tell anything about the distribution from any single data set. Generating multiple replications lets us see if the estimates are close, on average, to our assumption of constant ICC’s. Here is a function to generate a single data set:

genDD <- **function**(defc, defa, nclust, nperiods, waves, len, start) {

dc <- genData(nclust, defc)

dp <- addPeriods(dc, nperiods, "cluster")

dp <- trtStepWedge(dp, "cluster", nWaves = waves,

lenWaves = len, startPer = start)

dd <- genCluster(dp, cLevelVar = "timeID", "m", "id")

dd <- addColumns(defa, dd)

**return**(dd[])

}

And here is a function to estimate 200 sets of ICC’s for 200 data sets:

icc <- mclapply(1:200,

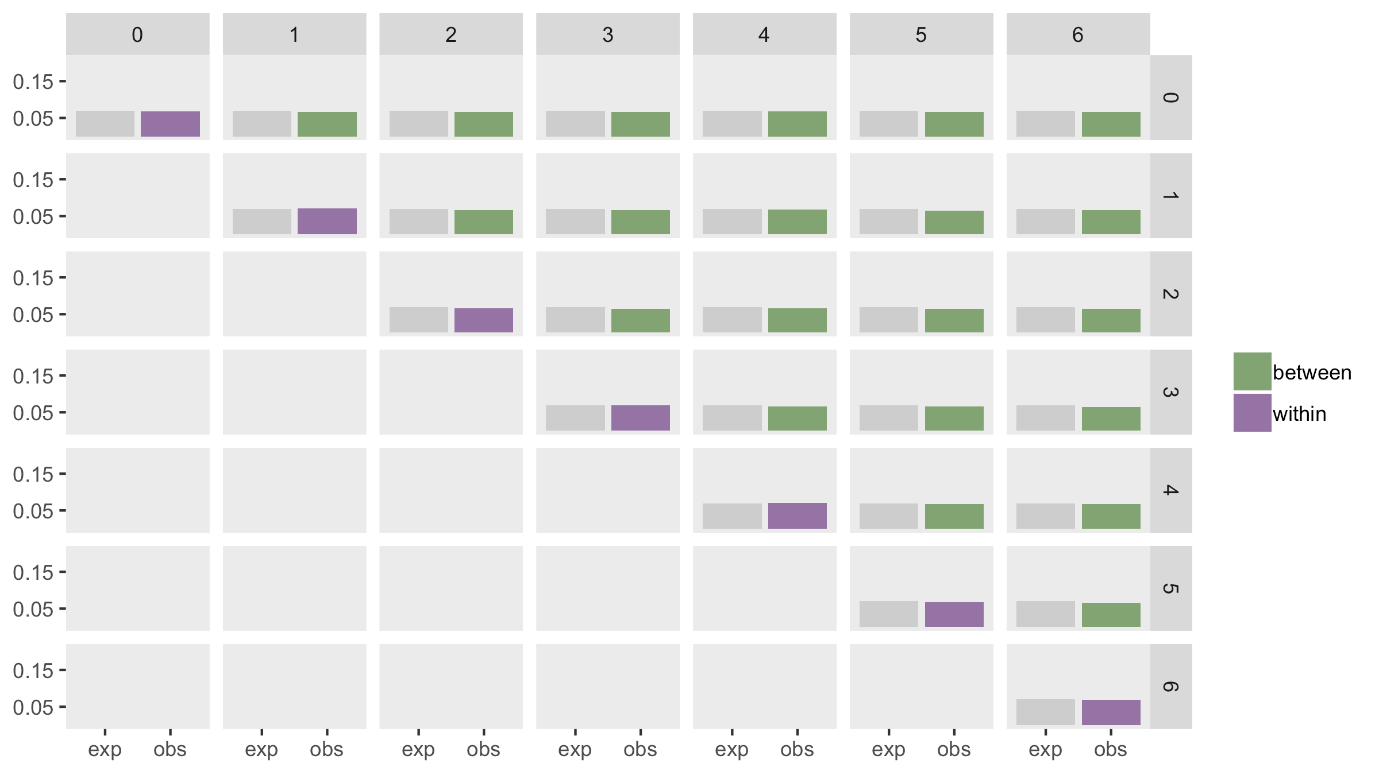
**function**(x) iccs(genDD(defc, defa, 100, 7, 4, 1, 2), byWave = T),

mc.cores = 4

)

observed <- sapply(rbindlist(icc), **function**(x) mean(x))

Averages of all the *within-* and *between-period* ICC’s were in fact quite close to the “true” value of 0.07 based on a relatively small number of replications. The plot shows the observed averages along side the expected value (shown in gray) for each of the periods generated in the data. There is little variation across both the *within-* and *between-period* ICC’s.



I’ll give you a little time to absorb this. Next time, I will consider alternative data generating processes where the the ICC’s are not necessarily constant.

**Different within- and between-period ICC’s**

In a scenario with constant within- and between-period ICC’s, the correlated data can be induced using a single cluster-level effect like \(b\_c\) in this model:

\[  
Y\_{ict} = \mu + \beta\_0t + \beta\_1X\_{ct} + b\_{c} + e\_{ict}  
\]

More complexity can be added if, instead of a single cluster level effect, we have a vector of correlated cluster/time specific effects \(\mathbf{b\_c}\). These cluster-specific random effects \((b\_{c1}, b\_{c2}, \ldots, b\_{cT})\) replace \(b\_c\), and the slightly modified data generating model is

\[  
Y\_{ict} = \mu + \beta\_0t + \beta\_1X\_{ct} + b\_{ct} + e\_{ict}  
\]

The vector \(\mathbf{b\_c}\) has a multivariate normal distribution \(N\_T(0, \sigma^2\_b \mathbf{R})\). This model assumes a common covariance structure across all clusters, \(\sigma^2\_b \mathbf{R}\), where the general version of \(\mathbf{R}\) is

\[  
\mathbf{R} =  
\left(  
\begin{matrix}  
1 & r\_{12} & r\_{13} & \cdots & r\_{1T} \\  
r\_{21} & 1 & r\_{23} & \cdots & r\_{2T} \\  
r\_{31} & r\_{32} & 1 & \cdots & r\_{3T} \\  
\vdots & \vdots & \vdots & \vdots & \vdots \\  
r\_{T1} & r\_{T2} & r\_{T3} & \cdots & 1  
\end{matrix}  
\right )  
\]

**Within-period cluster correlation**

The covariance of any two individuals \(i\) and \(j\) in the same cluster \(c\) and same period \(t\) is

\[  
\begin{aligned}  
cov(Y\_{ict}, Y\_{jct}) &= cor(\mu + \beta\_0t + \beta\_1X\_{ct} + b\_{ct} + e\_{ict}, \ \mu + \beta\_0t + \beta\_1X\_{ct} + b\_{ct} + e\_{jct}) \\  
\\  
&= cov(b\_{ct}, b\_{ct}) + cov(e\_{ict}, e\_{jct}) \\  
\\  
&=var(b\_{ct}) + 0 \\  
\\  
&= \sigma^2\_b r\_{tt} \\  
\\  
&= \sigma^2\_b \qquad \qquad \qquad \text{since } r\_{tt} = 1, \ \forall t \in \ ( 1, \ldots, T)  
\end{aligned}  
\]

And I showed in the previous post that \(var(Y\_{ict}) = var(Y\_{jct}) = \sigma^2\_b + \sigma^2\_e\), so the within-period intra-cluster correlation is what we saw last time:

\[ICC\_{tt} = \frac{\sigma^2\_b}{\sigma^2\_b+\sigma^2\_e}\]

**Between-period cluster correlation**

The covariance of any two individuals in the same cluster but two *different* time periods \(t\) and \(t^{\prime}\) is:

\[  
\begin{aligned}  
cov(Y\_{ict}, Y\_{jct^{\prime}}) &= cor(\mu + \beta\_0t + \beta\_1X\_{ct} + b\_{ct} + e\_{ict}, \ \mu + \beta\_0t + \beta\_1X\_{ct^{\prime}} + b\_{ct^{\prime}} + e\_{jct^{\prime}}) \\  
\\  
&= cov(b\_{ct}, b\_{ct^{\prime}}) + cov(e\_{ict}, e\_{jct^{\prime}}) \\  
\\  
&= \sigma^2\_br\_{tt^{\prime}}  
\end{aligned}  
\]

Based on this, the between-period intra-cluster correlation is

\[ ICC\_{tt^\prime} =\frac{\sigma^2\_b}{\sigma^2\_b+\sigma^2\_e} r\_{tt^{\prime}}\]

**Adding structure to matrix \(\mathbf{R}\)**

This paper by [*Kasza et al*](https://journals.sagepub.com/doi/full/10.1177/0962280217734981), which describes various stepped-wedge models, suggests a structured variation of \(\mathbf{R}\) that is a function of two parameters, \(r\_0\) and \(r\):

\[  
\mathbf{R} = \mathbf{R}(r\_0, r) =  
\left(  
\begin{matrix}  
1 & r\_0r & r\_0r^2 & \cdots & r\_0r^{T-1} \\  
r\_0r & 1 & r\_0 r & \cdots & r\_0 r^{T-2} \\  
r\_0r^2 & r\_0 r & 1 & \cdots & r\_0 r^{T-3} \\  
\vdots & \vdots & \vdots & \vdots & \vdots \\  
r\_0r^{T-1} & r\_0r^{T-2} & r\_0 r^{T-3} & \cdots & 1  
\end{matrix}  
\right )  
\]

How we specify \(r\_0\) and \(r\) reflects different assumptions about the between-period intra-cluster correlations. I describe two particular cases below.

**Constant correlation over time**

In this first case, the correlation between individuals in the same cluster but different time periods is less than the correlation between individuals in the same cluster and same time period. In other words, \(ICC\_{tt} \ne ICC\_{tt^\prime}\). However the between-period correlation is constant, or \(ICC\_{tt^\prime}\) are constant for all \(t\) and \(t^\prime\). We have these correlations when \(r\_0 = \rho\) and \(r = 1\), giving

\[  
\mathbf{R} = \mathbf{R}(\rho, 1) =  
\left(  
\begin{matrix}  
1 & \rho & \rho & \cdots & \rho \\  
\rho & 1 & \rho & \cdots & \rho \\  
\rho & \rho & 1 & \cdots & \rho \\  
\vdots & \vdots & \vdots & \vdots & \vdots \\  
\rho & \rho & \rho & \cdots & 1  
\end{matrix}  
\right )  
\]

To simulate under this scenario, I am setting \(\sigma\_b^2 = 0.15\), \(\sigma\_e^2 = 2.0\), and \(\rho = 0.6\). We would expect the following ICC’s:

\[  
\begin{aligned}  
ICC\_{tt} &= \frac{0.15}{0.15+2.00} = 0.0698 \\  
\\  
ICC\_{tt^\prime} &= \frac{0.15}{0.15+2.00}\times0.6 = 0.0419  
\end{aligned}  
\]

Here is the code to define and generate the data:

defc <- defData(varname = "mu", formula = 0,

dist = "nonrandom", id = "cluster")

defc <- defData(defc, "s2", formula = 0.15, dist = "nonrandom")

defa <- defDataAdd(varname = "Y",

formula = "0 + 0.10 \* period + 1 \* rx + cteffect",

variance = 2, dist = "normal")

dc <- genData(100, defc)

dp <- addPeriods(dc, 7, "cluster")

dp <- trtStepWedge(dp, "cluster", nWaves = 4, lenWaves = 1, startPer = 2)

dp <- addCorGen(dtOld = dp, nvars = 7, idvar = "cluster",

rho = 0.6, corstr = "cs", dist = "normal",

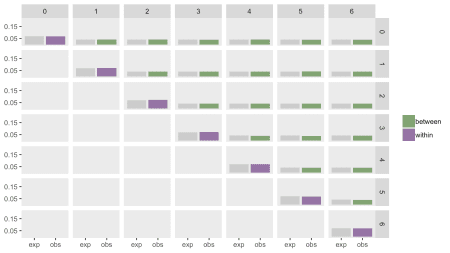
param1 = "mu", param2 = "s2", cnames = "cteffect")

dd <- genCluster(dp, cLevelVar = "timeID", numIndsVar = 100,

level1ID = "id")

dd <- addColumns(defa, dd)

As I did in my previous post, I’ve generated 200 data sets, estimated the *within-* and *between-period* ICC’s for each data set, and computed the average for each. The plot below shows the expected values in gray and the estimated values in purple and green.



**Declining correlation over time**

In this second case, we make an assumption that the correlation between individuals in the same cluster degrades over time. Here, the correlation between two individuals in adjacent time periods is stronger than the correlation between individuals in periods further apart. That is \(ICC\_{tt^\prime} > ICC\_{tt^{\prime\prime}}\) if \(|t^\prime – t| < |t^{\prime\prime} – t|\). This structure can be created by setting \(r\_0 = 1\) and \(r=\rho\), giving us an auto-regressive correlation matrix \(R\):

\[  
\mathbf{R} = \mathbf{R}(1, \rho) =  
\left(  
\begin{matrix}  
1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\  
\rho & 1 & \rho & \cdots & \rho^{T-2} \\  
\rho^2 & \rho & 1 & \cdots & \rho^{T-3} \\  
\vdots & \vdots & \vdots & \vdots & \vdots \\  
\rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1  
\end{matrix}  
\right )  
\]

I’ve generated data using the same variance assumptions as above. The only difference in this case is that the corstr argument in the call to addCorGen is “ar1” rather than “cs” (which was used above). Here are a few of the expected correlations:

\[  
\begin{aligned}  
ICC\_{t,t} &= \frac{0.15}{0.15+2.00} = 0.0698 \\  
\\  
ICC\_{t,t+1} &= \frac{0.15}{0.15+2.00}\times 0.6^{1} = 0.0419 \\  
\\  
ICC\_{t,t+2} &= \frac{0.15}{0.15+2.00}\times 0.6^{2} = 0.0251 \\  
\\  
\vdots  
\\  
ICC\_{t, t+6} &= \frac{0.15}{0.15+2.00}\times 0.6^{6} = 0.0032  
\end{aligned}  
\]

And here is the code:

defc <- defData(varname = "mu", formula = 0,

dist = "nonrandom", id = "cluster")

defc <- defData(defc, "s2", formula = 0.15, dist = "nonrandom")

defa <- defDataAdd(varname = "Y",

formula = "0 + 0.10 \* period + 1 \* rx + cteffect",

variance = 2, dist = "normal")

dc <- genData(100, defc)

dp <- addPeriods(dc, 7, "cluster")

dp <- trtStepWedge(dp, "cluster", nWaves = 4, lenWaves = 1, startPer = 2)

dp <- addCorGen(dtOld = dp, nvars = 7, idvar = "cluster",

rho = 0.6, corstr = "ar1", dist = "normal",

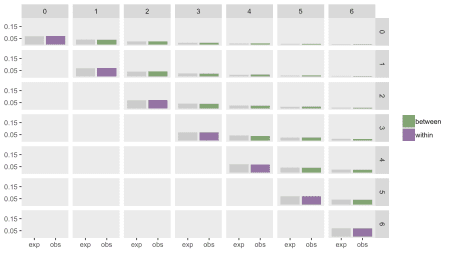
param1 = "mu", param2 = "s2", cnames = "cteffect")

dd <- genCluster(dp, cLevelVar = "timeID", numIndsVar = 10,

level1ID = "id")

dd <- addColumns(defa, dd)

And here are the observed average estimates (based on 200 datasets) alongside the expected values:



**Random slope**

In this last case, I am exploring what the ICC’s look like in the context of random effects model that includes a cluster-specific intercept \(b\_c\) and a cluster-specific slope \(s\_c\):

\[  
Y\_{ict} = \mu + \beta\_0 t + \beta\_1 X\_{ct} + b\_c + s\_c t + e\_{ict}  
\]

Both \(b\_c\) and \(s\_c\) are normally distributed with mean 0, and variances \(\sigma\_b^2\) and \(\sigma\_s^2\), respectively. (In this example \(\sigma\_b^2\) and \(\sigma\_s^2\) are uncorrelated, but that may not necessarily be the case.)

Because of the random slopes, the variance of the \(Y\)’s increase over time:

\[  
var(Y\_{ict}) = \sigma^2\_b + t^2 \sigma^2\_s + \sigma^2\_e  
\]

The same is true for the within- and between-period covariances:

\[  
\begin{aligned}  
cov(Y\_{ict}, Y\_{jct}) &= \sigma^2\_b + t^2 \sigma^2\_s \\  
\\  
cov(Y\_{ict}, Y\_{jct^\prime}) &= \sigma^2\_b + tt^\prime \sigma^2\_s \\  
\end{aligned}  
\]

The ICC’s that follow from these various variances and covariances are:

\[  
\begin{aligned}  
ITT\_{tt} &= \frac{\sigma^2\_b + t^2 \sigma^2\_s}{\sigma^2\_b + t^2 \sigma^2\_s + \sigma^2\_e}\\  
\\  
ITT\_{tt^\prime} & = \frac{\sigma^2\_b + tt^\prime \sigma^2\_s}{\left[(\sigma^2\_b + t^2 \sigma^2\_s + \sigma^2\_e)(\sigma^2\_b + {t^\prime}^2 \sigma^2\_s + \sigma^2\_e)\right]^\frac{1}{2}}  
\end{aligned}  
\]

In this example, \(\sigma^2\_s = 0.01\) (and the other variances remain as before), so

\[ ITT\_{33} = \frac{0.15 + 3^2 \times 0.01}{0.15 + 3^2 \times 0.01 + 2} =0.1071\]  
and

\[ ITT\_{36} = \frac{0.15 + 3 \times 6 \times 0.01}{\left[(0.15 + 3^2 \times 0.01 + 2)(0.15 + 6^2 \times 0.01 + 2)\right ]^\frac{1}{2}} =0.1392\]

Here’s the data generation:

defc <- defData(varname = "ceffect", formula = 0, variance = 0.15,

dist = "normal", id = "cluster")

defc <- defData(defc, "cteffect", formula = 0, variance = 0.01,

dist = "normal")

defa <- defDataAdd(varname = "Y",

formula = "0 + ceffect + 0.10 \* period + cteffect \* period + 1 \* rx",

variance = 2, dist = "normal")

dc <- genData(100, defc)

dp <- addPeriods(dc, 7, "cluster")

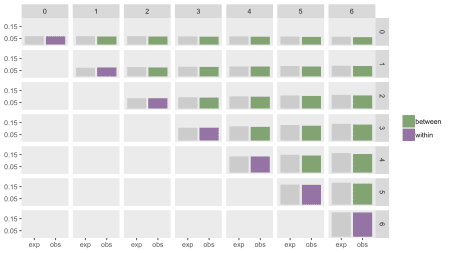
dp <- trtStepWedge(dp, "cluster", nWaves = 4, lenWaves = 1, startPer = 2)

dd <- genCluster(dp, cLevelVar = "timeID", numIndsVar = 10,

level1ID = "id")

dd <- addColumns(defa, dd)

And here is the comparison between observed and expected ICC’s. The estimates are quite variable, so there appears to be slight bias. However, if I generated more than 200 data sets, the mean would likely converge closer to the expected values.



In the next post (or two), I plan on providing some examples of fitting models to the data I’ve generated here. In some cases, fairly standard linear mixed effects models in R may be adequate, but in others, we may need to look elsewhere.

References:

Kasza, J., K. Hemming, R. Hooper, J. N. S. Matthews, and A. B. Forbes. “Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials.” *Statistical methods in medical research* (2017): 0962280217734981.