Theory

## Proportional-hazards models and intra-class correlation

Cox proportional-hazards models (or Cox regression) estimate the hazard of an event occurring in relation to predictor variables with time-to-event data (i.e., time taken till the event or censoring; Cox 1972). The hazard is a rate (or risk) of an event occurring at time . The hazard rate is defined in a Cox model as:

where is the hazard rate at time for the *i*th subject (individual), is the baseline hazard rate, are the regression coefficients, and are the predictor variables. Notably, takes the place of the intercept as or where (‘ln’ is a natural logarithm). Equation 1 can be rearranged to:

for the right hand side to take a linear form, which is more familiar for many readers although it does not have the intercept (i.e., ) and the residual term (i.e., ). To fit such a model using, for example, R, one needs to provide the time-to-event data in the form of a Surv object (e.g., Surv(time, event) where time is time taken till an event or censoring and event is usually 0 or 1, indicating whether the event occurred or not). The Cox model can be fitted using the coxph function in the survival package (Therneau et al. 2015).

Now let us assume that we have a single predictor variable, sex, for a time-to-event data set (e.g., a latency to solve a task), and we have a single random effect (or cluster) (e.g., individuals or populations). The Cox proportional-hazards model can be extended to include a random effect (individual identity), which is often referred to as the ‘frailty’ term and Cox regression with a single random effect is known as the frailty model:

where is the hazard rate at time for the *i*th individual for the *j*th occasion (observation). This model (frailty model) can be fitted using the coxme function in the R package coxme (Therneau 2015) as well as coxph.

Now we have defined the Cox model so let us define repeatably or intra-class correlations (ICC) in its simplest form when the trait of interest (the response variable) is a Gaussian variable (i.e., having normally distributed residuals):

where is the variance of the random effect (the between-cluster variance, where a cluster could be individual identity) and is the variance of the residuals (or within-cluster variance; Nakagawa & Schielzeth 2010). The ICC can be interpreted as the proportion of the total variance that is due to the between-cluster variance. The ICC can be calculated for (generalized) linear mixed-effect models (LMMS or GLMMs). For example, the R package, rptR can be used to calculate ICC from a variety of GLMMs, via the lmer and glmer function in the lme4 package.

Nakagawa & Schielzeth (2010) suggest that for non-Gaussian data (e.g. binomial or Poisson), the within-cluster variance can be determined by what distributional assumptions GLMM makes (e.g. binomial or Poisson). For example, to obtain ICC for binary GLMMs on the latent (link/transformed) scale, can be assumed to be (they call as the distributional specific variance; for more details, see Nakagawa & Schielzeth 2010. Nakagawa et al. 2017). However, Cox models do not make any distributional assumptions about the hazard rate (i.e., non-parametric; Equation 1 & 2). Although frailty models (Equation 3) have a random effect with a Gaussian distribution and so are referred to as semi-parametric. Therefore, using these current tools, we cannot calculate the ICC for Cox models.

Yet, in the statistical literature, a formula for the non-parametric version of ICC () for the frailty model is known when the random effect is disbursed as a Gamma distribution on the exponential scale. If we denote the variance from a Gamma distribution as under Equation 3, can be written as:

where the first and the second , are the shape and the rate parameter of the Gamma distribution, respectively (such parameterization results in the mean, and variance, ).

The estimate, is Kendall’s (the rank correlation or concordance for within-cluster observations) for the frailty model (Hougaard 2000). Unfortunately, there is no closed from (formula) when assuming a Gaussian distribution for the random effect as in Equation 3. Yet, can be obtained numerically and we provide an R function based on the tau function from the R package, parfm (Munda et al. 2012). We note that (Gaussian) and (Gamma) are unlikely to be the same. Yet, two values under two different assumptions (Gaussian and Gamma) are likely to be very similar, which we show in the simulation below.

An issue with the is that it is not a parametric version of ICC and more importantly, it is not clear whether this method can be extended where a Cox model has more than one random effect (at least, practically speaking). Therefore, we need to turn a time-to-event data set for Cox models into a data set where we could fit a GLMM to obtain parametric versions of ICC via GLMMs.

## Cox proportional-hazards models and generalized linear mixed models

In the statistical literature, it seems to be well known that the frailty model (Equation 3) can be fitted as a Poisson GLMM (known as, the piece-wise exponential model; e.g., Hirsh et al. 2016) or a binomial GLMM (the discrete-time model; Finkelstein 1986, Suresh et al. 2022; for an accessible account, see Austin 2017). Here, we show the discrete-time model, more specifically, the binomial GLMM with the complementary log-log (cloglog) link can be used to fit a comparable model as Equation 3 by “exploding” the time-to-event data through defining arbitrary discrete time intervals (Figure 1 shows an example of such an exploded data set compared to the original). If we assume we have three (arbitrary discrete) time intervals (*t1*, *t2* & *t3*) This binomial GLMM (without the intercept) is defined as:

where is the hazard rate at time for the *i*th subject for the *j*the occasions for the *k*th time interval (), is the baseline hazard rate for the th time interval, are the indicator variables for the time intervals, and are the regression coefficients for the time intervals. Note that the cloglog link is where is the probability of the event occurring so that the left hand side of Equation 6 consists of the cloglog-transformed hazard rate () and base hazard rate ().

Rather remarkably, and in Equation 6 are “the same” as those in Equation 3 although data structures used for two models are very different (i.e., time-to-event data vs. exploded data; Figure 1). Using a simple simulation, we show the equivalence of and between the Cox (frailty) model, fitted with coxph and coxme, and the binomial GLMM, fitted with glmer with event (0 or 1) as the response (see below). Note that the number of intervals does not affect these estimates from the binomial GLMM.

Therefore, we can use variance components obtained from Cox models to estimate ICC under a binomial GLMM with the complementary log-log link where (or the distributional-specific variance; Equation 4) is on the latent scale. This means we can define ICC for Equation 3 and 6 as (Nakagawa et al. 2017):

In the simulation below, we show the parametric version of ICC and the non-parametric version () are well correlated but not equivalent (analogous to the relationship between Pearson’s *r* and Kendall’s ). We prefer the use of ICC as in Equation 7 over because the parametric version is more comparable to other ICC estimates (on the latent/link scale) derived from GLMMs (e.g., Poisson and binomial data), which are now commonly used in ecology and evolution (Nakagawa et al. 2017).

Furthermore, the advantage of this approach is for us to add more than one random effect, as mentioned above. For example, imagine we have another cluster (a random effect such as population identity), adding it to Equation 3 yields:

where is the random effect for the th level of the second cluster, which is normally distributed with the mean of zero and the variance of (Cox models with more than one random factors cannot be fit with the coxph function, but can with coxme). It is interesting to notice that the two random effects can be ‘nested’ or ‘crossed’ (Schielzeth and Nakagawa, 2013).

An example of the nested random effects are individual () and population () where individuals are nested within populations. In this case, the ICC for individuals can be defined as:

The reason both variance components are required for ICC for individuals is that some of similarities of a pair of individuals come from belonging to the same population. An example of the crossed random effects are individual () and year () where individuals are not nested within years but observed across multiple years. This time, the ICC for individuals can be written as:

If one wants to remove or adjust for the effect of year, then ICC for individuals simplifies to Equation 7. Speaking of adjusting, all the ICC formulas presented above represent ‘adjusted’ repeatability (ICC; *sensu* Nakagawa and Schielzeth, 2010) because the effect of sex is accounted for in these models. We can obtain ‘unadjusted’ repeatablity (ICC) by fitting the model without the fixed effect (sex), for example by changing Equation 3:

Importantly, the and ICC values obtained from this model should be larger (Equation 11) than those obtained from Equation 3 given the fixed effect explains non-zero variance. In the supplementary materials (www.github….xxxx), we show how to fit models and obtain ICC estimates that we described above.

## Simulation results

Note - I do not think we need a section for simulation???

## Case studies