Examining an outcome (Y): mean and variance

Simple random sample (SRS)

Mean = Ybar = (1/N) * sum(Y1+Y2+Y3+...+YN)
Var(mean) = (1/N)^2 * [Var(Y1) + Var(Y2) + Var(Y3) +…+ Var(YN)]
Cov(Y1,Y2) = Cov(Yi,Yj) = 0

Cluster sample: For example, n subjects per cluster (or n timepoints per subject)

Mean = Ybar = (1/N) * sum(Y11+Y12+...+Y1n + Y21+Y22+...+Y2n + …)
Var(mean) = (1/N)^2 * [Var(Y11) + Var(Y12) + … + Var(Y1n) + … + 2*Cov(Y11,Y12) + …]
For Cov(Y1,Y2) > 0, etc, the variance under cluster sampling is LARGER than under SRS.

Paired, and examining difference pre to post (example via a slope):

Mean difference = Mean(Y2-Y1) (also = Mean(Y2) – Mean(Y1))
= (1/N) * sum((Ypost1-Ypre1) + (Ypost2-Ypre2) + … + (YpostN-YpreN))
Var(mean) = (1/N)^2 * [Var(Ypost1) + Var(Ypre1) – 2*Cov(Ypost1, Ypre1) + …]
For Cov(Ypost1,Ypre1) > 0, etc, variance under paired sampling is SMALLER than under SRS.

NOTE: Design effect – Due to the increased variance under cluster sampling, we need to enroll more than we would have needed under SRS (DE is the multiple).

Methods for correlated data

"**Fixed effects" (FE):** For a continuous outcome, this is like including a term for every group (cluster) as an explanatory variable. Otherwise can "condition out" the cluster effects (ex conditional logistic regression).

Random effects (RE): The effects for each cluster come from a random distribution (process), and this induces correlation within the cluster.

Generalized Estimating Equations (GEE): Separately specific structure for the mean model (incorporating any covariates), variances for each individual (ex each time point), and within-cluster associations (correlations).

For example, for the longitudinal model for subject i at times j:

CD4percent_ij = B0 + B1*time_j + B2*Ind[child <5months old at HAART initiation]_i + e_ij
[Plus any specifications needed to implement a FE, RE, or GEE approach.]

Thus we can consider longitudinal data as a type of correlated data, for which we are interested specifically in treating time as an explanatory variable in our model.

Matching – Once we account for the multiple units in a cluster, analysis follows as with any other repeated measures (correlated) data problem.

RE vs GEE

Interpretation: Cluster-specific vs Population-average (same if continuous outcome)
Missingness assumptions – MCAR vs MAR

Baseline response

Can include in the response vector (ie has its own row in the data)
OR subtract it from responses at follow-up time points ("change" analysis)
OR model the follow-up time points (or change) and ADJUST for the baseline value

Adjusting for baseline or not addresses different scientific questions: (1) whether individuals in each group have the same expected change over time; versus (2) whether individuals in each group have same expected change GIVEN THAT they have the same baseline response.

Generally could assume that groups have the same population means at baseline for randomized trials, but not for all observational studies.

STATA SYNTAX: FIGURES

Can use lowess to visualize the relationship between Y and X – point estimate.

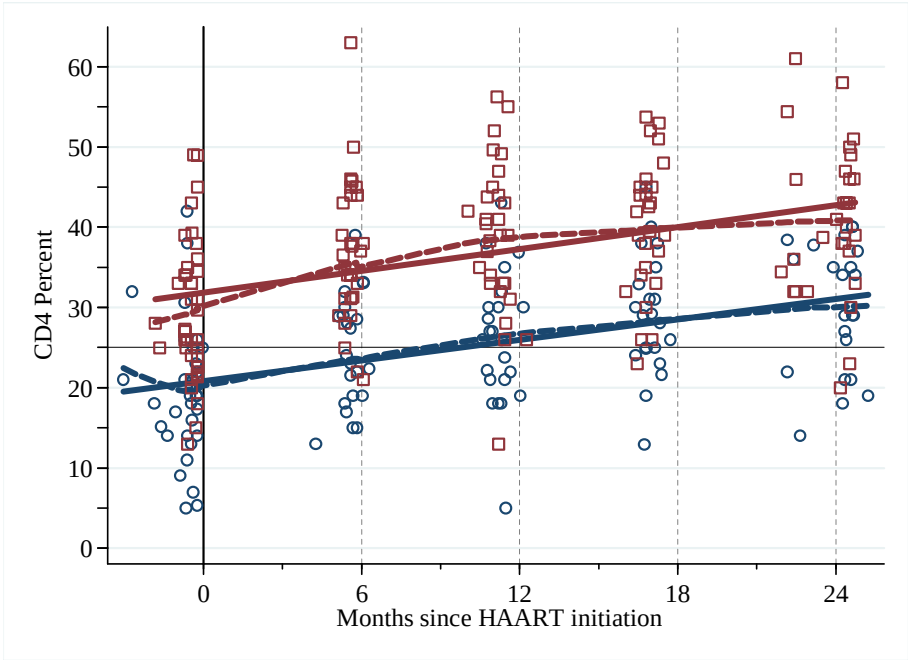
If linear fit is close to the lowess, could justify treatment of time as linear in the model.

Continuous (normally-distributed) and proportion outcomes

The model you fit will fully incorporate the correlations and give you final estimates for slopes, standard errors, p-values.

```
use "OutputData\Temp\main3.dta", clear
keep if ( HIdt~=. )

sort           ptid
merge 1:m      ptid using  "OutputData\Temp\longitudinal3.dta"
keep if ( _merge==3 )
drop _merge
sort ptid visitdate
keep if ( infcd4per ~=. )
keep if ( (visitdate <= randomdt) & (timehaart_months<25) )
twoway ///
       (scatter infcd4per timehaart_months if ( newgroup==0 ),  ///
              msymbol(O) mlcolor(navy)  mfcolor(none)   ///
              sort yaxis(1) ) ///
       (lowess  infcd4per timehaart_months if ( newgroup==0 ),  ///
              lpattern(dash) lwidth(thick) lcolor(navy) ) ///
       (lfit    infcd4per timehaart_months   if ( newgroup==0 ),  ///
              lpattern(solid) lwidth(thick) lcolor(navy) ) ///
       (scatter infcd4per timehaart_months if ( newgroup==1 ),  ///
              msymbol(S) mlcolor(maroon) mfcolor(none)   ///
              sort ) ///
       (lowess  infcd4per timehaart_months if ( newgroup==1 ),  ///
              lpattern(dash) lwidth(thick) lcolor(maroon) ) ///
       (lfit    infcd4per timehaart_months if ( newgroup==1 ),  ///
              lpattern(solid) lwidth(thick) lcolor(maroon) ) ///
       , ///
       ylabel(0(10)60, axis(1) angle(0))     ytick(0(5)60, axis(1)) ///
       yline(25.0, lwidth(vvthin) lcolor(black) ) ///
       ytitle("CD4 Percent", axis(1)) ///
       xscale(range(0 24)) xlabel(0(6)24) xmtick(0(3)24)  ///
       xline(0 6 12 18 24, lwidth(vvthin) lcolor(gray) lpattern(shortdash) ) ///
       xline(0.0, lwidth(medium) lcolor(black) ) ///
       xtitle("Months since HAART initiation") ///
       legend(off) ///
       graphregion(color(white)) scale(1.1)
```



Connected lines

```
twoway ///
(connect infcd4per timehaart_months if ( ptid=="0311001519" ),
   msymbol(O) mlcolor(navy) mfcolor(none)  lpattern(solid) lwidth(medthick) lcolor(navy)   sort yaxis(1) ) ///
(connect infcd4per timehaart_months if ( ptid=="0311002719" ),
   msymbol(O) mlcolor(navy) mfcolor(none)  lpattern(solid) lwidth(medthick) lcolor(green)  sort yaxis(1) ) ///
(connect infcd4per timehaart_months if ( ptid=="0311004119" ),
   msymbol(O) mlcolor(navy) mfcolor(none)  lpattern(solid) lwidth(medthick) lcolor(maroon) sort yaxis(1) ) ///
(connect infcd4per timehaart_months if ( ptid=="0311006319" ),
   msymbol(O) mlcolor(navy) mfcolor(none)  lpattern(solid) lwidth(medthick) lcolor(black)  sort yaxis(1) ) ///
(connect infcd4per timehaart_months if ( ptid=="0311009119" ),
   msymbol(O) mlcolor(navy) mfcolor(none)  lpattern(solid) lwidth(medthick) lcolor(orange) sort yaxis(1) ) ///
        , ///
        ylabel(0(10)60, axis(1) angle(0))     ytick(0(5)60, axis(1)) ///
        yline(25.0, lwidth(vvthin) lcolor(black) ) ///
        ytitle("CD4 Percent", axis(1)) ///
        xscale(range(0 24)) xlabel(0(6)24) xmtick(0(3)24)  ///
        xline(0 6 12 18 24, lwidth(vvthin) lcolor(gray) lpattern(shortdash) ) ///
        xline(0.0, lwidth(medium) lcolor(black) ) ///
        xtitle("Months since HAART initiation") ///
        legend(off) ///
        graphregion(color(white)) scale(1.1)
*graph export "C:\Users\ktapia\Figure2_CD4percent_Connected.eps", as(eps) fontface(Times) preview(off) replace
```