Sample Size Calculation for Longitudinal Studies

Phil Schumm

Department of Health Studies
University of Chicago

August 23, 2004

(Supported by National Institute on Aging grant P01 AG18911-01A1)

Make sample size calculations more accessible

- Permit researchers to experiment easily with many different scenarios
- Emphasize relationship between sample size calculation and actual analysis of the data

Longitudinal studies involve special considerations

- Correlation between data points for a given individual
- Sample attrition (drop-out)
- Effect(s) of interest often have complicated form (e.g., nonlinear across time)
- Effects of other covariates

Statistical model

Let $Y_i = (y_{i1}, \ldots, y_{in_i})^\mathsf{T}$ be an $n_i \times 1$ vector of outcome values for the ith subject ($i = 1, \ldots, m$), and $X_i = (x_{i1}, \ldots, x_{in_i})^\mathsf{T}$ be a corresponding $n_i \times p$ matrix of covariate values.

$$\begin{aligned} Y_i &= X_i\beta + \epsilon_i \\ \epsilon_i &\sim (0, V) \end{aligned}$$

Note:

- Within-subject correlation $R = V/\sigma^2$
- ϵ_i and ϵ_j are independent for all $i \neq j$

Generalized least squares estimator

$$\hat{\beta}_{GLS} = \left(\sum_{i=1}^{m} X_i' V^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{m} X_i' V^{-1} Y_i\right)$$

$$\begin{aligned}\text{Var}(\hat{\beta}_{GLS}) &= \left(\sum_{i=1}^{m} X_i' V_i^{-1} X_i\right)^{-1} \\ &= f(X, R, \sigma^2)\end{aligned}$$

Steps in using -xtgee- to calculate variance

1. Create a *pseudo-dataset* containing X and Y for m subjects
2. Enter the correlation matrix R
3. Use -xtgee- command with fixed correlation R and dispersion parameter σ^2

Time-averaged difference between two groups

Suppose:

- 2 groups of equal size ($m/2$ subjects each)
- All subjects measured at $n = 3$ time points (no drop-out)
- Constant within-subject correlation $\rho = 0.5$
- $\sigma^2 = 1$
- We want 90% power to detect a difference d of 0.25 at the two-sided 0.05 level.

$$\begin{aligned} m &= (4\sigma^2/nd^2)(1 + (n-1)\rho)(z_{\alpha/2} + z_\beta)^2 \\ &= 448 \text{ (224 per group)} \end{aligned}$$

Create pseudo-dataset and enter correlation matrix

```
. input g t

             g           t
  1. 0 1
  2. 0 2
  3. 0 3
  4. 1 1
  5. 1 2
  6. 1 3
  7. end

. expand 224
(1338 observations created)

. bys t (g): gen id = _n

. gen y = uniform()

. mat R = J(3,3,0.5) + I(3)*(1-0.5)

. mat list R

symmetric R[3,3]
    c1  c2  c3
r1   1
r2  .5   1
r3  .5  .5   1
```

Use -xtgee- to calculate variance

```
. xtgee y g, i(id) t(t) corr(fixed R) scale(1)

Iteration 1: tolerance = 2.308e-16

GEE population-averaged model                 Number of obs       =      1344
Group and time vars:                 id t     Number of groups    =       448
Link:                            identity     Obs per group: min  =         3
Family:                          Gaussian                    avg  =       3.0
Correlation:             fixed (specified)                   max  =         3
                                              Wald chi2(1)        =      0.01
Scale parameter:                        1     Prob > chi2         =    0.9266

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           g |  -.0071089   .0771517    -0.09   0.927    -.1583235    .1441056
       _cons |   .5041997   .0545545     9.24   0.000     .3972749    .6111245
------------------------------------------------------------------------------
```

Use -sampsi- to calculate power

```
. sampsi 0 0.25, alpha(0.05) n1(1) sd1(0.0771517) onesample

Estimated power for one-sample comparison of mean
  to hypothesized value

Test Ho: m =      0, where m is the mean in the population

Assumptions:

         alpha =   0.0500  (two-sided)
 alternative m =      .25
            sd =  .077152
 sample size n =        1

Estimated power:

         power =   0.8998
```

Difference in rate of change between two groups

- 10% drop-out per time point
- Constant within-subject correlation $\rho = 0.8$
- Linear change in each group over time
- We want to compute power for detecting a 0.25 difference in the rate of change over entire study.

Create pseudo-dataset and enter correlation matrix

```
. xtdes, i(id) t(t)

      id:  1, 2, ..., 448                                    n =        448
       t:  1, 2, ..., 3                                      T =          3
           Delta(t) = 1; (3-1)+1 = 3
           (id*t uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       1       3         3         3       3       3

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
      364     81.25   81.25 |  111
       44      9.82   91.07 |  1..
       40      8.93  100.00 |  11.
 ---------------------------+---------
      448    100.00         |  XXX

. mat R = J(3,3,0.8) + I(3)*(1-0.8)
```

Use -xtgee- to calculate variance

```
. xi: xtgee y i.g*t, i(id) t(t) corr(fixed R) scale(1)
i.g               _Ig_0-1             (naturally coded; _Ig_0 omitted)
i.g*t             _IgXt_#             (coded as above)

Iteration 1: tolerance = .00579064
Iteration 2: tolerance = 4.388e-16

GEE population-averaged model                 Number of obs        =      1216
Group and time vars:                    id t   Number of groups     =       448
Link:                               identity   Obs per group: min =         1
Family:                             Gaussian                  avg =       2.7
Correlation:                fixed (specified)                 max =         3
                                               Wald chi2(3)         =      0.18
Scale parameter:                           1   Prob > chi2          =    0.9808

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _Ig_1 |  -.0121577    .1075186    -0.11   0.910    -.2228903    .1985749
           t |    .007212    .0229825     0.31   0.754    -.0378328    .0522568
     _IgXt_1 |   -.002222    .0325021    -0.07   0.945     -.065925     .061481
       _cons |   .5004684    .0760271     6.58   0.000      .351458    .6494789
------------------------------------------------------------------------------
```

Use -sampsi- to calculate power

```
. loc d = 0.25/3

. sampsi 0 `d', alpha(0.05) n1(1) sd1(0.0325021) onesample

Estimated power for one-sample comparison of mean
  to hypothesized value

Test Ho: m =      0, where m is the mean in the population

Assumptions:

         alpha =   0.0500  (two-sided)
 alternative m =  .083333
            sd =  .032502
 sample size n =        1

Estimated power:

         power =   0.7271
```

Extensions of the basic method

- Generalized linear models

$$\mathrm{Var}(\hat{\beta}) = f(X, \beta, R, \phi)$$

- Effects of other covariates

  Note: For random X, may need to increase size of pseudo-dataset and then scale up variance estimate accordingly.