Title

aldsim - Data Simulation for Accelerated Longitudinal Designs (ALDs)

Version

Code Version 1.0 on 09/29/2017 programmed in Stata Version 15.0 by Nicholas J. Jackson

Syntax

aldsim, nc(number) cn(number) ci(number) pn(number) pi(number) [options]

options Description

-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Design Options

nc(number) Number of people per cohort. Alternately can specify totn.

totn(number) Total sample size. nc=[totn/cn].

cn(number) Number of cohorts.

ci(number) The interval between cohorts in years. May default to pi if not specified.

pn(number) Number of periods.

pi(number) The interval between periods in years.

Age Options

agestart(number) Age at 1st period for youngest cohort (starting age).

agesd(number) Standard deviation of agestart. Used for when ages are generated from Normal or Truncated Normal distribution.

agelb(number) Lower Bound for agestart. Specifies minimum age for the 1st period of youngest cohort. Used for when ages are generated from Truncated Normal or Uniform distribution. See Options for more information.

ageub(number) Upper Bound for agestart. Specifies maximun age for the 1st period of youngest cohort. Used for when ages are generated from Truncated Normal or Uniform distribution. See Options for more information.

Slope and Intercept Options

slp(number) Age slope.

slpsd(number) Standard deviation of the age slope (a random slope).

int(number) Intercept value.

intsd(number) Standard deviation of the intercept (a random intercept).

slpintcor(number) Correlation between random intercept and slope.

Non-Linear Growth Options

nltype(string) Type of non-linear model. Options are log for logistic-normal (default), exp for exponential growth, or gom for Gompertz.

alpha(number) Rate of approach to the upper asymptote. Treated as a fixed parameter unless alphasd is specified.

alphasd(number) Standard deviation of the alpha parameter. Specification turns these additive non-linear models into multiplicative non-linear models.

ageinflect(number) Age at which the rate of change is greatest. By default this value is constant between cohorts and specified as the median of the ages under study.

ageinflectshift(number)

Difference between cohorts in the age at which the rate of change is the greatest.

Effect Size Options

effsize(number) Effect Size. Defined as effsize = slp/slpsd. Default=0.5.

gcr(number) Growth curve reliability. Proportion of variance explained by growth parameters. Can be specified instead of resid option. Default=0.8.

resid(number) Standard deviation of the residuals. Can be specified in place of gcr.

Cohort Differences Options

cid(number) Cohort Intercept Differences. Difference in intercept between sucessive cohorts.

cidratio(number) Cohort Intercept Difference Ratio. Ratio of cid to intsd.

csd(number) Cohort Slope Differences. Difference in slope between sucessive cohorts.

csdratio(number) Cohort Slope Difference Ratio. Ratio of csd to slpsd.

fixedcohort Specifies fixed cohort differences (as opposed to random). See Options for more information.

Period Differences Options

period(number) Slope for fixed effect of period.

periodshift(number) Difference in period slope between sucessive cohorts.

Attrition Options (attrition, gamma, and attrtype must be specified together)

attrition(numlist) Amount of study attrition (w). Proportion of participants who drop out at some point. Details on methods for attrition found in Remarks.

gamma(numlist) Specifies where drop-out should be concentrated. γ>1, concentrated towards the end. γ<1 dropout occurs more often in the beginning.

attrtype(string) Specifies if attrition applied to the ALD should be based on the study length in the ALD or in the Single Cohort Design (SCD). Options are ald or scd. Details on choosing a method for attrition is in

Options below.

attrstatic Specifies how attrition probabilities should be applied when choosing which observations to remove.

Output Options

graph Produces a figure of the simulated data.

print Displays design information on-screen.

-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Description

aldsim simulates data from an accelerated longitudinal design (ALD) for linear and non-linear models with and without attrition as well as with options for cohort and period effects. Data from this program can then be

analyzed using standard mixed model approaches. This program is part of a larger suite of programs for designing, simulating, and estimating accelerated longitudinal designs. Please see aldesign, aldcost, and aldest for

more information.

Acronyms

ALD - Accelerated Longitudinal Design.

SCD - Single Cohort Design. Refers to the traditional longitudinal design comprised of a single cohort covering the same age span as the ALD.

Options

+----------------+

----+ Design Options +-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

nc(number); The number of subjects per cohort may be specified as integer values. This program assumes an equal N per cohort such that the total sample size is defined as Nc\*Cn. For those with designs where this assumtption

is not correct (i.e. uneqal N per cohort) they should modify Nc so that Nc\*Cn is equal to the total sample size they will have. Alternately used may specify the totn option which provides the total sample size and

computes Nc = totn/Cn.

cn(number); The number of cohorts may be specified as integer values ≥1.

ci(number); The interval spacing (difference in age) between two adjacent cohorts (in years) may be specified as values ≥0.

pn(number); The number of periods may be specified as integer values ≥1.

pi(number); The interval spacing (difference in age) between two adjacent periods (in years) may be specified as values ≥0.

+-------------+

----+ Age Options +----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

agestart(number); The mean age (in years) at the first period for the youngest cohort. When specified with agesd, a normal distribution for ages at each period is created. When specified with agesd and agelb and ageub, a

truncated normal distribution is created.

agesd(number); The standard deviation of the age distribution at the first period for the youngest cohort. When specified with agestart, a normal distribution for ages at each period is created. When specified with agestart

and agelb and ageub, a truncated normal distribution is created.

agelb(number); Lower bound (absolute minimum value) for the age distribution at the first period for the youngest cohort. Must be specified with ageub. When specified without agestart, agestart = (ageub - agelb)/2. In

absence of agesd, the age distribution is a uniform distribution over ages agelb to ageub. When specified with agesd and ageub, a truncated normal distribution is created.

ageub(number); Upper bound (absolute maximum value) for the age distribution at the first period for the youngest cohort. Must be specified with agelb. When specified without agestart, agestart = (ageub - agelb)/2. In

absence of agesd, the age distribution is a uniform distribution over ages agelb to ageub. When specified with agesd and ageub, a truncated normal distribution is created.

+-----------------------------+

----+ Slope and Intercept Options +------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

slp(number); The value for the slope of age. Created as a random slope when slpsd is specified.

slpsd(number); The standard deviation of the slope for age (within-cohort variation). Allows the slope to be specified as a random variable. Slope can alternately be specified as fixed by setting slpsd=0.

int(number); The value for the intercept. Created as a random intercept when intsd is specified.

intsd(number); The standard deviation of the intercept (within-cohort variation). Allows the intercept to be specified as a random variable. The intercept can alternately be specified as fixed by setting intsd=0. Defaults to

the value of slpsd if not specified.

slpintcor(number); Correlation between the slope and intercept. Set to 0 if unspecified.

+---------------------------+

----+ Non-Linear Growth Options +--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

nltype(string); The type of non-linear model to use. Options are:

log Log-Normal Growth (default): y= int + slp/(1+exp(-alpha\*(age-ageinflect)))

exp Exponential Growth: y= int + slp\*(1-exp(-alpha\*(age-ageinflect)))

gom Gompertz Function: y= int + slp\*exp(-exp(-alpha\*(age-ageinflect)))

alpha(number); The rate of approach for the non-linear growth. When alphasd is not specified, alpha is treated as fixed.

alphasd(number); The standard deviation for the rate of approach. When alphasd is specified, alpha is treated as random normal variable and the nonlinear models are considered multiplicative.

ageinflect(number); The age at which the rate of change is the greatest. Age in the non-linear equations will be centered around this value. When not specified; ageinflect = (agestart + pi'\*(pn-1) + ci\*(cn-1))/2

ageinflectshift(number); Specifies the between cohort differences (fixed effects) in the ageinflect. Assumes linear effects for the age-inflection shift such that the age at which the rate of change is greatest for Cohort n

will be: ageinflect\* = ageinflect + ageinflectshift\*(n-1)

+---------------------+

----+ Effect Size Options +--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

effsize(number); The effect size for the slope, defined as: effsize = slp / slpsd. When not specified, default=0.5. The recommended use of effsize is to specify it in conjunction with slp and thereby avoid needing to

specify the slpsd value.

gcr(number); The growth curve reliability. When not specified, default=0.8. The GCR is defined as the proportion of variability explained by the growth parameters and is computed by: gcr = (intsd^2 + (agestart^2 \* slpsd^2) +

(2\*agestart\*slpintcor\*intsd\*slpsd)) / (intsd^2 + (agestart^2 \* slpsd^2) + (2\*agestart\*slpintcor\*intsd\*slpsd) + resid^2). The recommended use of gcr is to specify it instead of resid. Please see Hertzdog, Lindenberger,

Ghisletta, and von Oertzen (2006) for more information on the GCR.

resid(number); Standard deviation of the residuals distributed as a random normal variable. When gcr is specified, resid will be computed from the gcr.

+----------------------------+

----+ Cohort Differences Options +-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

cid(number); The Cohort Intercept Difference is defined as the difference between sucessive cohorts in their intercept value. When specified, the intercept for Cohort n is: int\* = int + cid\*(n-1). As is apparent from the

equation, the difference between cohorts is treated linearly. By default, cid is drawn from the positive values of a random normal distribution of mean 0 and sd=cid. That is: cid\*=abs(rnormal(0, cid)). As a result, the

actual intercept difference between two sucessive cohorts may be greater or less than the cid specified. If the intercept differences are desired to be exactly cid, users should specify the fixedcohort option.

cidratio(number); The ratio of between-cohort to within-cohort variance in intercept. Defined as: cidratio = cid / intsd Users can specify cid with cidratio in place of intsd.

csd(number); The Cohort Slope Difference is defined as the difference between sucessive cohorts in their slope values. When specified, the age slope for Cohort n is: slp\* = slp + csd\*(n-1). As is apparent from the equation,

the difference between cohorts is treated linearly. By default, csd is drawn from the positive values of a random normal distribution of mean 0 and sd=csd. That is: csd\*=abs(rnormal(0, csd)). As a result, the actual

slope difference between two sucessive cohorts may be greater or less than the csd specified. If the slope differences are desired to be exactly csd, users should specify the fixedcohort option.

csdratio(number); The ratio of between-cohort to within-cohort variance in slope. Defined as: csdratio = csd / slpsd Users can specify csd with csdratio in place of slpsd.

fixedcohort; Specifies that cid and csd be treated as fixed values as opposed to being drawn from a random normal distribution. When specified the difference bettween two sucessive cohorts in their intercept and slope values

will be exactly cid and csd.

+----------------------------+

----+ Period Differences Options +-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

period(number); For linear models only. Allows a fixed linear period effect of the nature period\*(period-1) that is constant across cohorts.

periodshift(number); Used in conjunction with period. Allows for fixed cohort differences in the period effect such that the period slope for Cohort n is defined as: period\* = period + periodshift\*(n-1).

+-------------------+

----+ Attrition Options +----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Note: The Attrition Options of attrition, gamma, and attrtype must be specified together.

attrition(numlist); Specifies the attrition percent, such that by the end of the study this % of subjects will have missing data. Values must be between 0 and 1. This value is called w in attrition equation in the Remarks

section below.

gamma(numlist); Specifies where drop-out should be concentrated. Values must be > 0. Values > 1 indicate drop-out concentrated towards the end of the study. Values < 1 indicate drop-out occurs more often in the beginning of

the study. This value is called γ in attrition equation in the Remarks section below.

attrtype(string); Indicates how attrition in the ALD should be applied. Values allowed are scd or ald. When scd is specified, the attrition in the ALD is calculated based on the values at each measurement in the SCD. For

example if there will be 25% attrition by the end of 10 measurements in the SCD, an ALD with only 4 measurements will have it's attrition values calculated at the 4th measurement in the SCD. This will result in the ALD

having less than 25% drop-out by the end of the 4th measurement. This has the effect of making the ALD a much better choice relative to the SCD when attrition is applied. Using scd is the recommended choice for most

instances as this assumes that drop-out is related to overall time in the study rather than being specific to the study design. Choosing ald will result in attrition values in the ALD being based on the number of

measurements in the ALD. For example, the attrition in an SCD would still be calculated in the same manner, such that with 25% attrition there will be 25% of subjects with some drop-out at the end of 10 measurements.

The attrition for an ALD with only 4 measurements however would now also be caluclated so that there is 25% attrition by the end of the 4th measurement. This implies that both the SCD and ALD should have the same

drop-out rates despite the SCD being a longer study. However, this option might be desirable under circumstances where the interest is not comparing values to the SCD.

attrstatic; Removal of observations based on the attrition probabilities (see Remarks below), by default is based on comparison to a random uniform distribution. In practice, this means that from one simulation to the next,

the number of subjects with exactly j measurements will vary. If instead the user would like the number of subjects with exactly j measurements to remain the same (static) from one simulation to the next, the attrstatic

should be specified. attrstatic will compare the attrition probabilies to the value of \_n/\_N for each cohort, which will result in the same number of observations being removed between different simulations with the same

design parameters.

Remarks

Development of the models

The models employed here are based on the age-period-cohort mixed model specification for aggregate data from O'Brien, Hudson, and Stockard (2008). By adapting their model to incorporate within-person repeated measurement,

the ALD Mixed Model can be created.

y\_ijk = b0 + b1\*age + u0\_j + u0\_k + u1\_j\*age + u1\_k\*age + e\_ijk

Where b0 is the fixed intercept, b1 the fixed slope, u0\_j the within-cohort between-person variability in intercept (random intercept), u0\_k the between-cohort variability in intercept (random intercept), u1\_j the

within-cohort between person variability in slope (random coefficient), u1\_k the between-cohort variability in slope (random coefficient), and e\_ijk the residuals. Fixed (or random) effects for period could also be specified

in this general description of the ALD Mixed Model (ALDMM).

Incorporating Attrition

Taking the approach of Galbraith, Bowden, & Mander (2017) that was originally proposed by Verbeke & Lesaffre (1999), we will utilize a Weibull model for defining the probabilities of attrition. Each participant has 1 to M

measurements, for which each measurement can be defined by the probability of dropping out of the study p = {p\_1, ... p\_j ... , p\_M}, where p\_j is the probability of having exactly j measurements. The probabilities in p are

then determined based on the Weibull function. The proportion of individuals who have dropped out at period t, assuming they remained in the study until period t, will be defined by:

λγt^(γ-1), where λ = -log(1-w) and w is the amount of attrition and gamma (γ) the drop-out concentration.

The periods of measurement can be rescaled as proportions from 0 to 1 by using the following formula:

t\_j\*=(t\_j-1)/(M-1)

The first period (t\_1) is equal to zero and the last (t\_M) is equal to 1. Implied in this rescaling is that the period intervals are equivalent across all periods.

Using these rescaled periods, we can compute the proportion of subjects who drop-out after measurment t\_j by:

p\_j = (1 - w)^(t\_j^γ) - (1 - w)^(t\_j+1^γ), where p\_M = 1 - w

In this manner, for a given value of w, the dropout serves as a function of γ whereby when γ=1, the dropout is constant over the course of study. When γ>1, dropout is concentrated towards the end of the study and when γ<1

dropout occurs more often in the beginning of the study. When γ=1, missing data are presumed missing completely at random (MCAR) and when γ≠1 data are presumed missing at random (MAR).

-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Examples

Linear model using a truncated normal age distribution. Default gcr and effsize values are used. Simulation output paramters printed and graphed. No between cohort differences in intercept or slope.

. aldsim, nc(100) cn(5) ci(2) pi(2) pn(8) agestart(10) agesd(0.5) agelb(9.5) ageub(10.5) print graph

Linear model using a truncated normal age distribution. gcr=0.7 and effsize=3 with slp=10. Simulation output paramters printed and graphed. No between cohort differences in intercept or slope.

. aldsim, nc(100) cn(5) ci(2) pi(2) pn(8) agestart(10) agesd(0.5) agelb(9.5) ageub(10.5) gcr(0.7) effsize(3) slp(10) print graph

Linear model using a truncated normal age distribution. gcr=0.7 and effsize=3 with slp=10. Simulation output paramters printed and graphed. Between cohort variance in slope is 4 times greater than within-cohort slope SD. Not

between-cohort differences in intercept. Cohort differences are specified as fixed parameters (i.e. fixedcohort option).

. aldsim, nc(100) cn(5) ci(2) pi(2) pn(8) agestart(10) agesd(0.5) agelb(9.5) ageub(10.5) gcr(0.7) effsize(3) slp(10) csdratio(4) fixedcohort print graph

References

Galbraith, S., Bowden, J., & Mander, A. (2017). Accelerated longitudinal designs: an overview of modelling, power, costs and handling missing data. Statistical Methods in Medical Research, 26(1), 374-398.

Hertzog, C., Lindenberger, U., Ghisletta, P., & Oertzen, T. V. (2006). On the power of multivariate latent growth curve models to detect correlated change. Psychological Methods, 11(3), 244.

O'Brien, R. M., Hudson, K., & Stockard, J. (2008). A mixed model estimation of age, period, and cohort effects. Sociological Methods & Research, 36(3), 402-428.

Verbeke, G., & Lesaffre, E. (1999). The Effect of Drop‐Out on the Efficiency of Longitudinal Experiments. Journal of the Royal Statistical Society: Series C (Applied Statistics), 48(3), 363-375.