Duke GLOBAL HEALTH INSTITUTE | Research Design & Analysis Core

Core Guide: Longitudinal Data Analysis

Part of a series addressing common issues in statistical and epidemiological design and analysis

Background

In contrast to *cross-sectional* data, which are collected at a single time point, longitudinal data are collected at multiple time points on the same individuals over time. These so called *repeated measures* data may be related to an exposure, or an outcome/event, or both. The primary benefit of collecting longitudinal data is the ability to prospectively record the health outcome, as well as to measure an exposure that may be associated with this outcome. Longitudinal studies are generally considered superior to cross-sectional studies in validly estimating risk; and prospectively measuring an exposure will reduce the possibility of misclassification of this exposure that frequently occurs in retrospective studies. Furthermore, longitudinal data allow for measurement of changes in outcomes over time within a single unit of analysis (e.g., an individual), and can tease apart different types of time-dependent effects; namely Age, Period and Cohort effects.

Box A. Differences Between Three Common Structures of Longitudinal data

Longitudinal cohort data usually include a short time series of repeated measured on the same unit of analysis (e.g. individuals). This usually consists of many units of analysis with a few repeated observations within units. Analysis of longitudinal cohort data are the focus of this guide.

Repeated cross-sectional data consist of repeated measures on different individuals, or other unit of analysis, over time. Since repeated measures are on different individuals, we do not expect correlation among these measures due to individual characteristics; but we may observe correlation due to external factors, such as environmental or other seasonal effects.

Time series data usually consist of a longer time series on a single (or small number of) individuals, or other unit of analysis. With time series analysis, we assume temporal correlation in the response, but assume that if observations are far enough apart, they are essentially independent. In analyzing time series data, we are often focused on making inference to temporal dynamics within a population, such as the pattern and intensity of dengue fever, and less interested in inference to individual level risk factors for disease.

This guide will review the most common longitudinal study designs, as well as the most appropriate methods for analyzing the resulting data. Since longitudinal data consist of repeated, and thus, *correlated*, measures on the same individual, or other unit, such as a village, appropriate analyses must be considered when analyzing data that exhibit this correlation structure. This guide will summarize analytic techniques for handling response correlation and will provide example Stata and SAS analysis code. For a more detailed, technical discussion of modeling the correlation structure of longitudinal data, see the DGHI Core Guide titled, ***Correlation Structures in Longitudinal Data Analysis*.**

Worked Example

For the remainder of this guide, we will consider a longitudinal study of HIV-positive pregnant women in rural Uganda. The investigators want to determine whether exposure to a new intervention reduces HIV viral load among study participants. To this end, the investigators collect data on the outcome Y (i.e., viral load), exposure A (i.e., exposed to the intervention), as well as any other variables of interest, such as time period, T, or other covariates, L. Furthermore, the outcome is measured for all study participants at baseline and at least one follow-up time point (i.e. it is a longitudinal cohort study). This guide will consider several longitudinal study designs that may be used to answer this research question. We will review each of the primary designs questions listed in the table below, and discuss their implications for our resulting analyses.

No.	Question	(Yes/No)
1a	Is there a 'control' group in the study – i.e., participants who were not exposed to the intervention?	
1b	⟹ If yes, then was the exposure randomized across participants; i.e., is the study experimental?	
2	Were the same participants measured at each time point?	
3	Is there more than one follow-up time point?	

Pre-test/Post-test – no controls

The pre-test/post-test ('pre-post') design is one of simplest forms of longitudinal studies (Figure 1). The study often consists of a single baseline measurement, $Y_{t=0}$, and is compared to a single follow-up measurement, $Y_{t=1,}$ usually occurring after an intervention. Using our worked example, we might wish to measure viral load among HIV-positive pregnant women in a rural health facility in Uganda, before and after an intervention is initiated aimed at reducing viral load.

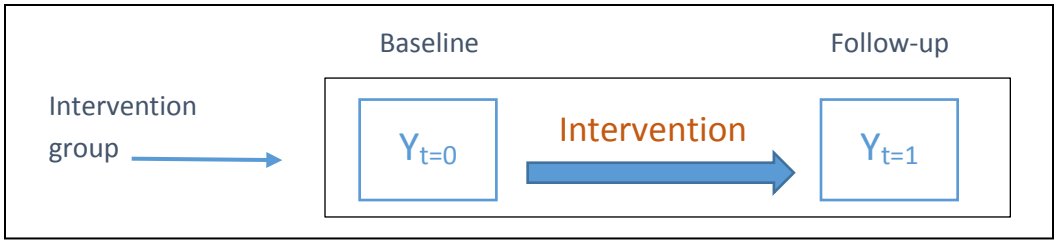

Figure 1. Schematic depiction of a pre-post study design, no controls.

Since there is no control group (i.e.., participants who were unexposed to the intervention), the baseline measurement is our best estimate of the counterfactual; that is, what would be the value of viral load for our study participants at t=1, had they remained unexposed to the intervention. In this design, we answer our above key questions in the following way:

No.	Question	(Yes/No)
1a	Is there a 'control' group in the study – i.e., participants who were not exposed to the intervention?	No
1b	⟹ If yes, then was the exposure randomized across participants; i.e., is the study experimental?	n/a
2	Were the same participants measured at each time point?	Yes
3	Is there more than one follow-up time point?	No

Statistical analysis

Our pre-post data could be analyzed by taking the difference in the baseline and follow-up measurements and analyzing the resulting data. For example, if our outcome is viral load (i.e. a continuous variable, which we assume to follow a Normal distribution), we might test the null hypothesis that the mean change in viral load over time is equal to zero. We would difference the two values for each study participant, and then run a **one-sample T-test** to determine whether the observed mean CD4 count is different from zero (*See **Appendix** for Stata and SAS code for performing all statistical analyses described in this document*). Doing so would eliminate any correlation in the response since the investigator would have already removed it by differencing the two values, i.e. this analytic method is an appropriate method to account for the positive correlation that is expected between two outcomes measured on the same individual.

Alternatively, we could retain both the baseline and follow-up measurements and perform a **paired T-test** to determine whether, on average, the difference in the two values is significantly different from zero. This paired analysis will give us the exact same result (T-statistic) as the one-sample T-test above.

In the event that our outcome is binary (e.g., 0/1), we could perform a paired test of categorical data, such as a **McNemar's test** to test the null hypothesis: is the proportion of participants with the event different, comparing the baseline and follow-up time points? For example, in our pre-post – no control

design in HIV+ women, we might be interested in whether the proportion of those women with viral suppression (yes/no) is different at baseline and at follow-up. For this analysis, viral load suppression would be categorized as 0 or 1, instead of using viral load as a continuous measure as we did for our T-test.

Pre-test/Post-test - with controls

Two important sources of bias that can confound our effect estimate (e.g., mean difference in viral load) in a pre-post study are: 1.*individual-level covariates;* and 2.*time*. In the pre-post study design described above, we are able to deal with individual-level time-fixed covariates (e.g. gender) as a potential confounder, since each person acts as their own control. However, we are not able to account for the potential confounding effects of time. To illustrate this, let's imagine that during the same time period that we performed our intervention to lower viral load among HIV-positive pregnant mothers, there was a national advertising campaign aimed at educating HIV-positive Ugandans on the importance of adherence to antiretroviral therapy (ART). If we saw a significant decrease in viral load in our study participants over the study period, we would not be able to estimate whether this change was due to our study intervention, or whether it was due to the national advertising campaign, or some combination of both. Said another way, we would not be able to adjust our effect measure for the potential confounding effects of time that, in this case, could be due to the advertising campaign. **Adding a control group of participants who are not exposed to the intervention of interest**, and who are followed over the same time period as our intervention participants, is likely the best method for reducing the potential confounding effect of time (Figure 2). We note that some disciplines classify the "pre-test/post-test – with controls" design as a "quasi-experimental" design. For reasons described in Box B, we do not use this term to describe any of the designs outlined in the current guide.

Box B. Use of the term "Quasi-experimental"

In some research disciplines, the term *quasi-experimental* refers broadly to observational studies where the exposure of interest is not randomized, and can include pre-post studies with and without controls. Some authors limit use of the term to only studies with a control group, while still others, particularly in health economics, use the term primarily to refer to *natural experiments*, where through some 'natural' mechanism, the exposure is pseudo-randomized in the target population. Epidemiologists do not typically use this term and prefer to distinguish observational studies by the specific design (e.g., case-control, cohort, regression discontinuity, etc.). To avoid further confusion, we do not use the term *quasi-experimental* in this document.

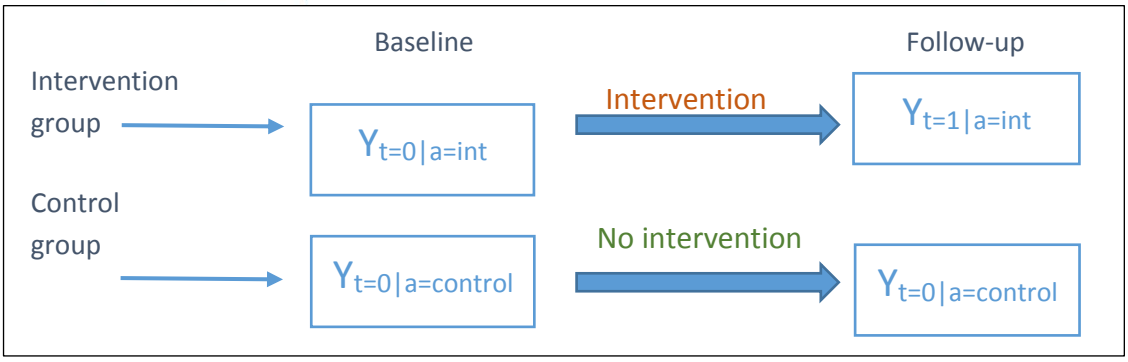

Figure 2. Schematic depiction of a pre-post study design, with controls

In this design, we answer our key study questions in the following way:

No.	Question	(Yes/No)
1a	Is there a 'control' group in the study – i.e., participants who were not exposed to the intervention?	Yes
1b	⟹ If yes, then was the exposure randomized across participants; i.e., is the study experimental?	No
2	Were the same participants measured at each time point?	Yes
3	Is there more than one follow-up time point?	No

Statistical analysis

So how would our analysis plan differ if we add a control group to our study? Just as we did before, we could analyze these data by taking the difference between the pre-post measurements. But instead of performing a paired T-test, to determine if the mean pre-post change in response (e.g., viral load) is zero, we would perform an unmatched **two-sample Student's T-Test** comparing the mean pre-post change in response over time between the control and intervention groups. This is called a *difference-in-difference analysis*. First, we difference the baseline and follow-up measures for each individual ($Y_{t=0} - Y_{t=1}$). Second, we compare the mean difference for intervention and control participants ($\bar{Y}_{control} - \bar{Y}_{int}$) by using a two-sample T-test. Since our first differencing is done on the same person, we implicitly account for the baseline response for each individual. However, the T-test is unable account for any confounding effects, like different age distributions in the two groups, that may be at least partly responsible for a difference in mean response across the two groups. In order to account for confounding, we would need to model the data using a method, such as linear regression[1].

Van Belle et al. note several ways to model these pre-post longitudinal data using regression methods, *including: 1. follow-up only*; 2. *change analysis*; and 3. analysis of covariance (*ANCOVA)*[1].

1. **Follow-up only** models only the follow-up response, ignoring baseline.

[1] Generalized linear modeling can be used with very similar parameterization when the response variable is non-Normal (e.g., binary, count).

$$Y_{i1} = B_0 + B_1X_i + \varepsilon_i, \quad (1)$$

where in our study Y_{i1} is the viral load measure at follow-up (i.e. time =1) for the ith study participant, X_i is an indicator variable denoting intervention group and ε_i is the error term. In this model, B_0 is the model estimated mean viral load at follow-up for the control group and B_1 is is the difference in the mean response at follow-up comparing $X_i = 1$ (e.g., intervention group) to $X_i = 0$ (e.g., control group). This unadjusted regression model is equivalent to a two-sample T-test comparing follow-up measures only (i.e., it ignores the baseline measure of the response). If the assignment to X_i (0,1) were randomized, then the simple follow-up comparison is a valid causal analysis of the effect of the treatment, and should be very similar to our crude estimate from the two-sample T-test. Otherwise, additional parameters representing individual-level characteristics, such as age or gender, may be added to the model to reduce bias due to confounding. Of the three models described in the current section, model (1) would also be the only choice in the situation where no baseline measurements of the outcome were available.

2. Change analysis models the difference in outcome from baseline to follow-up. First, the researcher would difference the two response values for each participant ($Y_{i1} - Y_{i0}$), and then regress this single value on X_i, i.e. the same indicator of the exposure group as specified in the **follow-up only model (1)**. The **change analysis** model would take the form

$$(Y_{i1} - Y_{i0}) = B_0 + B_1X_i + \varepsilon_i. \quad (2)$$

However, in the **change analysis model**, B_1 is interpreted as the difference between the mean change in response for intervention (i.e. $X_i = 1$) as compared to the average change in response for control (i.e. $X_i = 0$). This unadjusted regression model is equivalent to a two-sample T-test comparing the difference in baseline and follow-up measures.

3. Analysis of covariance (**ANCOVA)** models the difference in mean follow-up of the outcome, adjusting for baseline value of the outcome, and takes the form,

$$Y_{i1} = B_0 + B_1X_i + B_2Y_{i0} + \varepsilon_i. \quad (3)$$

Instead of differencing out the baseline value before fitting the regression model as was the case with the change analysis (2), in ANCOVA, we include the baseline value (Y_{i0}) as a predictor of follow-up. In ANCOVA, B_1 is the effect for the difference in mean follow-up outcome comparing intervention vs. control groups as it was in the follow-up only analysis (1), with the important difference that it is conditional on the other predictors in the model, namely the baseline level of the outcome. In this case, B_2 represents the effect of the baseline measurement. ANCOVA analysis should only be used when the exposure is randomized, such as in a randomized controlled trial (RCT). Van Belle et al. note that when the intervention is randomized, then B_1 from each of these three models (i.e., follow-up, change analysis, ANCOVA) provides identical estimates, meaning that each of these betas provides a valid estimate of the average causal effect of the intervention. However, ANCOVA has been shown to provide more precise effect estimates under certain conditions[2] (Chapter 5, pp 124-125).

Longitudinal cohort studies with multiple repeated time points

In cases where longitudinal data include more than two repeated measures[2], the researcher has the ability to estimate changes in the response over time, and assess whether this change varies by exposure. The underlying hypothesis may be the same as with a pre-post analysis (i.e., does the response vary over time by exposure status?), but a regression model with more than two repeated measures, may allow for a more nuanced understanding of this change over time, as well as increase the power and precision of your estimates. Figure 3 depicts a study with a baseline response measure and three follow-up measures in both a control group and an intervention group, but the design can be generalized to any number of follow-up time points and may not involve a control group.

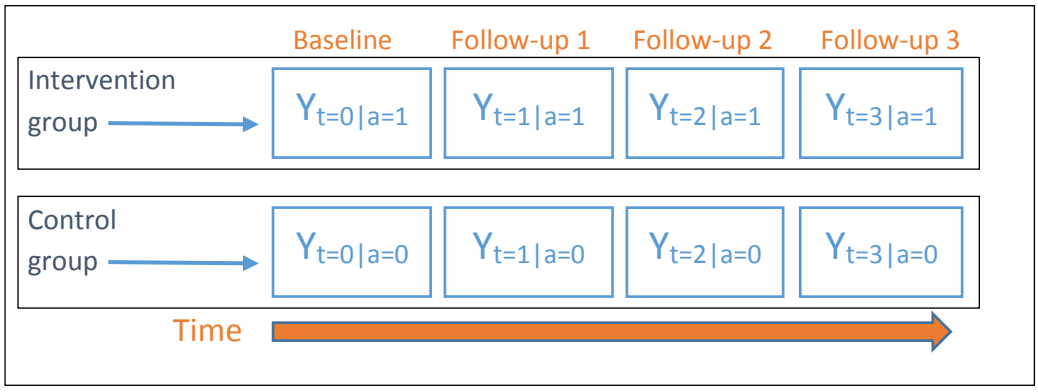

Figure 3. Schematic depiction of a repeated measures longitudinal study, with controls, with a baseline response measurement and three follow-up response measures.

In the longitudinal cohort design (Figure 3), we answer our key study questions for a non-randomized design in the following way:

No.	Question	(Yes/No)
1a	Is there a 'control' group in the study – i.e., participants who were not exposed to the intervention?	Yes
1b	⟹ If yes, then was the exposure randomized across participants; i.e., is the study experimental?	No
2	Were the same participants measured at each time point?	Yes
3	Is there more than one follow-up time point?	Yes

[2] We limit our discussion to cases where the follow-up time is fixed (e.g., 6 months) and measurements are taken at the same intervals for all individuals. Analyses that account for different follow-up times for study participants is beyond the scope of this Guide.

Statistical analysis

The statistical analysis plan for a multiple repeated measures longitudinal study has several important considerations, over and above what we have already discussed; and these considerations may depend on the research question being asked, as well as how the data were generated.

When our exposure is non-randomized (Q1b = No), then we would usually allow the mean response at baseline to vary between our exposure groups and we would fit indicator variables for each of the modeled time points, as well as their interaction with exposure level. If we performed an observational study with four repeated measures of the response (baseline + 3 follow-up), the regression model might take the form:

$$Y_{it} = B_0 + B_1 X_i + B_2(Time1_{it}) + B_3(Time2_{it}) + B_4(Time3_{it}) + B_5(X_i * Time1_{it}) + B_6(X_i * Time2_{it}) + B_7(X_i * Time3_{it}) + u_i + \varepsilon_{it}, \quad (4)$$

where B_0 is the mean response at baseline for the control group, B_1 is the mean response at baseline for the intervention group, B_2 - B_4 are the mean CHANGE in response values for the control group at the three follow-up times compared to baseline, respectively, and B_5 - B_7 are the CHANGE in mean response values (from baseline) in intervention participants compared to the CHANGE in mean response in control participants at the three follow-up times, respectively. In this case, because we anticipate correlation of responses measured on the same person, we also include a random intercept term u_i for each individual, where u_i is assumed to be normally distributed with zero-mean and a variance σ_u^2 to be estimated using the data. This model is called a *linear mixed effects model* and the residual error term, ε_{it}, is now different for each measurement on each person but is still assumed to be normally distributed with mean zero and constant variance, σ_e^2. Additional details and methods to account for this correlated error structure can be found in the section accounting for correlated errors.

We could also choose to perform a **change analysis**, just as we did with our pre-post data. This would require us to difference the baseline value from each of the three follow-up time points and regress this differenced measure. The regression model would be similar to our first method but now we have one fewer time point since it was used in the differencing step.

$$(Y_{it} - Y_{i0}) = B_0 + B_1 X_i + B_2(Time2_{it}) + B_3(Time3_{it}) + B_4(X_i * Time2_{it}) + B_5(X_i * Time3_{it}) + u_i + \varepsilon_{i,} \quad (5)$$

where B_0 is the mean CHANGE (from baseline) in response at Time 1 for the control group, B_1 is the difference in mean CHANGE in response at Time 1 for the intervention group compared to the control group, B_2 and B_3 are the mean CHANGE in response values for the control group at follow-up times 2 and 3, respectively, and B_4 and B_5 are the difference in CHANGE in mean response values (from baseline) in intervention participants compared to the CHANGE in mean response in control participants at follow-up times 2 and 3, respectively.

Now let's suppose that exposure was randomized across study groups. In this case, we would answer our key study questions in the following way (i.e., Q1b now = Yes):

No.	Question	(Yes/No)
1a	Is there a 'control' group in the study – i.e., participants who were not exposed to the intervention?	Yes
1b	⟹ If yes, then was the exposure randomized across participants; i.e., is the study experimental?	Yes
2	Were the same participants measured at each time point?	Yes
3	Is there more than one follow-up time point?	Yes

We may choose to fit a reduced form of model (4) without allowing the baseline mean response to vary by group. Although with randomized experiments it is also common practice to include this additional parameter as we do in Equation 4, it is not necessarily needed if groups have similar baseline values, as we would expect with randomization. When model (4) is modified to have a common baseline level, it is sometimes referred to in the literature as *constrained longitudinal data analysis* (cLDA).[3, 4] This model removes the fixed effect for 'Group' and would instead take the form:

$$Y_{it} = B_0 + B_1(B_1 X_i) + B_2(\text{Time2}_{it}) + B_3(\text{Time3}_{it}) + B_4(X_i*\text{Time1}_{it}) + B_5(X_i*\text{Time2}_{it} + B_6(X_i*\text{Time3}_{it}) + u_i + \varepsilon_{i,} \quad (6)$$

And finally, we could also perform an ANCOVA, as we did with our pre-post data. Using our multiple time points data, our equation would take the form

$$Y_{it} = B_0 + B_1(Y_{i0}) + B_2(X_i) + B_3(\text{Time2}_{it}) + B_4(\text{Time3}_{it}) + B_5(X_i*\text{Time2}_{it}) + B_6(X_i*\text{Time3}_{it}) + u_i + \varepsilon_{i,} \quad (7)$$

where B_0 is now the mean response for the control group at <u>time=1</u>, Y_{i0} is the response value for the i^{th} individual at baseline and B_2 is the mean response for the intervention group at time=1. In their book, *Applied Longitudinal Data Analysis*, Fitzmaurice, Laird and Ware note that results from fitting equations 6 and 7 yield almost identical results when the exposure is randomized[2]. So why would one choose the cLDA approach (equation 6) over ANCOVA (Equation 7)? The primary reason is that the cLDA model is often more efficient (i.e., more powerful) than the ANCOVA approach.[4] In addition, the cLDA model uses data from all subjects and all time points, even those who were only assessed at baseline. Fitzmaurice, Laird, and Ware also note that equations 6 and 7 should <u>not be used</u> when the exposure is non-randomized (i.e., observational studies)[2]. Equation 6 should not be used in this instance because it's often unreasonable to assume that mean baseline values would be equal across the two groups in an observational study. Furthermore, Equation 7 (i.e., ANCOVA) should not be used for observational studies because the baseline value may be associated with the exposure. To illustrate this last point, imagine that we fit equation 6 to our Ugandan HIV study where exposure to the intervention was non-

randomized. And let's further assume that because exposure was not randomized, people who were sicker (i.e., higher viral load) at baseline, were more likely to receive the intervention that those less sick[3]. In this case, if we fit equation 6 to the data, we would be testing the hypothesis,

> "Is the change in viral load over time different comparing a study participant who was exposed to the intervention and a participant who was unexposed"?

If, instead, we fit equation 7 to the data, we would be testing a different hypothesis; specifically,

> "Is the change in viral load over time different comparing a study participant who was exposed to the intervention and a participant who was unexposed, given they both have the same initial viral load value"?

This conditional hypothesis is rarely of interest in an observational study, since there is no reason to think that these two groups would have the same initial viral load value. If the study were randomized, however, we would expect that the groups would have similar mean baseline viral load values, as well as similar distributions of all other potential confounding factors.

Accounting for correlated errors

When modeling longitudinal repeated measures data, our usual assumption that each response is independent does not hold. This is because, by design, we have collected multiple responses on a single person or other unit of analysis, and these repeated measures are more similar to one another, on average, than they are to responses from another individual. If we do not properly account for this correlation in the response, we would underestimate the amount of variation in the response, which would lead to an underestimate of our modeled standard errors. There are three common methods for handling this correlated error structure: calculating **robust standard errors**, modeling the error by using a **generalized estimating equation,** and including a **random intercept** in a mixed effects regression model, i.e., as described above in equations (4)-(7). A detailed description of each method is beyond the scope of this Guide, but a brief overview and comparison of each, is described here. For a more detailed, technical discussion of modeling the correlation structure of longitudinal data within a mixed effects regression framework, see the Core Guide titled, ***Correlation Structures in Longitudinal Data Analysis*.**

Estimating **robust standard errors** simply means that we calculate the variance of the response directly from the sample, and we make no assumptions about the distribution (e.g., Gaussian) from which the data were generated. The most common method of generating robust standard errors is by way of the Hubert White sandwich estimator.

Using a **generalized estimating equations (GEE**) approach is a bit more complicated, but similar to calculating robust standard errors. The approach does not directly model the correlation structure (i.e.,

[3] This lead to a bias known as 'confounding by indication'.

the response covariance matrix), but rather treats it as a nuisance parameter. The difference is that you must specify a working correlation matrix. The four most common covariance structures are Independence, Exchangeable, Unstructured and Autoregressive. Although you must choose one as your working correlation matrix, under certain assumptions, GEE is robust to misspecification of this working correlation structure. The GEE approach can be used with any GLM, such as linear, logistic or log regression.

Fitting a **random intercept** in a mixed effects regression model is the only method of the three that explicitly models the response covariance matrix. In our repeated measures longitudinal design, we would fit a random intercept for study participant, which would model a separate 'random' parameter for the baseline response (e.g., viral load). Most commonly, this random intercept is assumed to be distributed Gaussian, in effect, allowing each participant's baseline response to vary from the mean across all participants. Modeling this random intercept allows a partitioning of the variance into *within* and *between* participant variance, which can be used to calculate the correlation, or 'clustering' of the response between each participant. The relative magnitude of the between-person variance to the overall variance of the outcome, i.e. $\sigma_u^2/(\sigma_u^2 + \sigma_e^2)$ is called the *intraclass correlation coefficient* (ICC). It is a useful statistic on its own, and can be used in the design of future studies.

Different meanings of "fixed effect"

In **biostatistics**, the term *fixed effect* refers to effects in a regression equation that are assumed not to vary; i.e., they are constant across individuals. This is in contrast to *random effects*, which are assumed to vary across individuals.

In **econometrics and other social sciences fields**, the term *fixed effect* means something quite different. It refers to a regression technique used for panel data *(panel data* is the econometrics term for longitudinal data) whereby the response, Y_{it}, is averaged across time, $\bar{Y}_{i,}$ and then this average value is differenced out of the regression equation. This technique is used to remove the autocorrelation found in repeated measures to achieve a similar goal as we have in biostatistics when we use GEE or GLMM.

In practice, fitting a random intercept implies an *exchangeable* correlation structure, which means that all pairs of outcome measurements on the same individual are assumed to have the same correlation irrespective of how far apart in time they are measured. Other random terms, e.g., a random slope, can also be included in the model in order to allow for alternative correlation structures. See [1, 3] for more details.

Marginal versus Conditional Effects: When to use GEE versus a Mixed Effects Regression model?

Fitting generalized linear models with either a GEE or within the mixed effects regression framework (e.g., with a random intercept) have become common practice when analyzing repeated measures longitudinal data. And while many researchers often use these two methods interchangeably, there are a couple of important considerations when determining which method to use for your study.

First, the parameter estimates from a linear GEE model with exchangeable working correlation matrix should be identical to a linear mixed model with a random intercept; however, **this is not true for non-linear models**. For example, the logit link used in logistic regression models is a non-linear transformation, which means that the mean for the random effect is estimated on the logit scale and will differ from the GEE estimate when back-transformed.

Second, and perhaps most importantly, the interpretation of results from a GEE model is different than that of a mixed model, even if the parameter estimates are identical. A GEE model is a so called, *population average* model. Inference is made to the *marginal* (i.e., the average effect across the entire population) of the exposure on the outcome. A mixed model, on the other hand, produces subject-specific (i.e., conditional) effects. For example, in the case of a mixed model with a random intercept, inference is conditional on the random intercept (i.e. the "subject"). Let's use our Uganda HIV study to illustrate this distinction. In this study, our main interest is to estimate the effect of the intervention on viral load level. Since we are analyzing repeated measures on the same individual, we need to use either a GEE model or a mixed model to account for this correlation in the response within study participants. If we choose a GEE model, we would interpret our parameter for 'group' as the mean difference in viral load comparing our intervention to our control group, *averaged* across the population of all individuals in the study. If we choose a mixed model instead, we would interpret our parameter for 'group' as the mean difference in viral load comparing our intervention to our control group, *conditional* on each subject's random effects. In this example, since viral load is modeled as a continuous outcome assuming a Gaussian distribution, the mixed model estimates can also be interpreted as population average. But again, the interpretation of the two models would not be the same for other GLM, such as logistic regression models.

Summary

This Core Guide has focused on basic analytic methods and considerations for common structures of data that arise from longitudinal cohort studies. The key design questions and corresponding analysis options are summarized in Table 1 and both Stata and SAS code to implement the options are provided in the Appendix.

Prepared by: Joseph Egger, PhD
Reviewed by: Duke Global Health Institute | Research Design & Analysis Core
Version: 1.0, last updated October 05, 2017

References

1. Van Belle, G. and L. Fisher, *Biostatistics : a methodology for the health sciences*. 2nd ed. Wiley series in probability and statistics. 2004, Hoboken, NJ: John Wiley & Sons. xi, 871 p.
2. Fitzmaurice, G.M., N.M. Laird, and J.H. Ware, *Applied longitudinal analysis*. 2nd ed. Wiley series in probability and statistics. 2011, Hoboken, N.J.: Wiley. xxv, 701 p.
3. Liang, K.-Y. and S.L. Zeger, *Longitudinal data analysis of continuous and discrete responses for pre-post designs.* Sankhyā: The Indian Journal of Statistics, Series B, 2000: p. 134-148.
4. Liu, G.F., et al., *Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials?* Statistics in medicine, 2009. **28**(20): p. 2509-2530.

Table 1. Summary table of common methods for analyzing longitudinal data.

Design	Control group?	Randomized?	Same pp measured?	More than 1 f/up?	Common analysis options
Designs with no control group					
Pre-post no control group (different people at each time point)	No	No	No	No	Distr~N: Two-sample Student's T-Test Distr~Binary: Chi-squared test of independence *Note: both methods are subject to individual-level confounding since populations differ at baseline and f/up.
Pre-post no control group (same people at each time point)	No	No	Yes	No	Distr~N: One-sample T-test; Paired T-test Distr~Binary: McNemar's test (paired) *Note: both methods are subject to confounding by time.
Designs with a control group – all with same individuals measured at each time point					
Pre-post with control group – not randomized	Yes	No	Yes	No	Distr~N: Two-sample Student's T-Test (on change in mean response) Distr~Binary: Chi-squared test of independence (at f/up only) *Note: both methods are subject to confounding since exposure is not randomized. **Regression methods** (assuming proper adjustment for confounding) - Follow-up only (Equation 1) - Change analysis (Equation 2)
Pre-post with control group –	Yes	Yes	Yes	No	Distr~N: Two-sample Student's T-Test (on change in mean response).

randomized (i.e. RCT)					Distr~Binary: Chi-squared test of independence (at f/up only). **Regression methods** - Follow-up only (Equation 1) - Change analysis (Equation 2) - ANCOVA (Equation 3)
	Yes	No	Yes	Yes	- Model full response vector with parameter for group (Equation 4) - Change analysis (Equation 5)
	Yes	Yes	Yes	Yes	- Model full response vector but excluding parameter for group (Equation 6) - ANCOVA (Equation 7)

Sample Stata and SAS Statistical Code

Code is provided here for example use only. The precise code needed for your analysis will likely differ.

1. One sample t-test (requires wide format data)

Stata

```
generate y1y0_diff = y1-y0
ttest y1y0_diff ==0
```

SAS

```
proc tttest data = name;
     var y;
run;
```

2. Paired t-test (requires wide format data)

Stata

```
ttest y0==y1
```

SAS

```
proc ttest data = name;
      paired y0*y1;
run;
```

3. McNemar's test (requires wide format data)

Stata

```
mcc y_bl y_fup [fw=n]
```

SAS

```
proc freq data=data;
     tables y_bl*y_fup /agree expected norow nocol nopercent;
run;
```

4. Two sample T-test

Stata

```
ttest y, by(groupvar) /*long format data*/

ttest y1 == y2, unpaired /*wide format data*/
```

SAS (assumes long format data)

```
proc ttest data = name;
     class(groupvar);
     var y;
run;
```

5. Chi-squared test for independence

Stata

```
tabulate y group, chi2
```

SAS (assumes long format data)

```
proc freq data = name;
     tables y*group / expected chisq;
run;
```

Regression models – all models below assume continuous Normal data but can be adapted to GLM. The options code for common GLMs with a binary response are:

Stata

```
Logistic: family(binomial) link(logit) eform

Log binary: family(binomial) link(log) eform

Linear risk: family(binomial) link(identity)
```

SAS

```
Logistic: dist = binomial link=logit;
estimate 'Beta' xvarname 1 -1/ exp;

Log: dist = binomial link=log;
estimate 'Beta' xvarname 1 -1/ exp;

Linear risk: dist = binomial link=identity
```

6. Regression: follow-up only (Equation 1)

Stata

```
glm y1 i.group, fam(gaussian) link(identity) /*assumes wide data.
fitting MLE. Use 'regress' command to fit equivalent model using OLS*/
```

```
glm y i.group if time==1, fam(gaussian) link(identity) /*assumes long
data where there is a single response variable, y, and a separate
indicator variable, time, that indicates baseline (0) or f/up (1)
period.
```

SAS (assumes long format data)

```
proc genmod data = name if time=1;
      class group;
      model y = group / dist = normal
                        link = identity;

run;
```

7. Regression: change analysis (Equation 2, assumes wide data)

Stata

```
generate y1y0_diff = y1-y0

glm y1y0_diff i.group, fam(gaussian) link(identity)
```

SAS

```
data data2;
      set data 1;
      y1y0_diff = y1-y0;
run;
proc genmod data = data2;
      class group;
      model y1y0_diff = group / dist = normal
                                link = identity;
run;
```

8. Regression: ANCOVA (Equation 3, assumes wide data)

Stata

```
glm y1 i.group y0, fam(gaussian) link(identity)
```

SAS

```
proc genmod data = data1;
      class group;
      model y1 = group y0 / dist = normal
                            link = identity;
```

```
run;
```

9. Regression: model full response vector with parameter for group (Equation 4, assumes long data)

All of the models below use either GEE or GLMM to account for correlated errors in the response, but assume a simple structure. See Core Guide *Correlation Structures in Longitudinal Data Analysis* for a more detailed description and sample code for these analyses.

Stata

GEE model

```
xtgee y i.group i.time i.group#i.time, fam(gaussian) link(identity)
i(id) corr(exc) eform /*correlation can be change to various options.
Most common are exchangeable, independent, unstructured and
autoregressive 1 (ar1)*/

estat wcorr /*provides the working correlation matrix from model*/
```

GLMM using mixed command

```
mixed y i.group i.time i.group#i.time || id: /*'mixed is the command
for linear mixed models. See 'help me' in stata for similar commands
for GLMM*/
```

SAS

GEE model

```
proc genmod data=data ;
   class group time id;
   model y = group time group*time  /  dist=normal;
   repeated subject=id / type=exch covb corrw;
run;
```

GLMM

```
proc mixed data=data;
      class group time id;
      model y = group time group*time / s chisq;
      random INTERCEPT / subject=id;
run;
```

10. Regression: change analysis (Equation 5, assumes long data)

Stata

This first step requires the baseline response, y0, to be stored as a separate variable. The data are in long form, however, there is no row for the baseline response. If there is, you will need to exclude from the analysis on the model command line ('if y !=0')

```
generate ydiff = y-y0
```

GEE model

```
xtgee ydiff i.group i.time i.group#i.time, fam(gaussian)
link(identity) i(id) corr(exc) eform
```

GLMM

```
mixed ydiff i.group i.time i.group#i.time || id:
```

SAS

GEE model

```
proc genmod data=data ;
   class group time id;
   model ydiff = group time group*time  /  dist=normal;
   repeated subject=id / type=exch covb corrw;
run;
```

GLMM

```
proc mixed data=data;
   class group time id;
   model ydiff = group time group*time / s chisq;
   random INTERCEPT / subject=id;
run;
```

11. Regression: Model full response vector but excluding parameter for group (Equation 6, assumes long data)

Stata

Data are assumed to be similar long form as with equation 4 above.

GEE model

```
xtgee y i.time i.group#i.time, fam(gaussian) link(identity) i(id)
corr(exc) eform
```

GLMM

```
mixed y i.time i.group#i.time || id:
```

SAS

GEE model

```
proc genmod data=data ;
   class group time id;
   model  y = time group*time  /  dist=normal;
   repeated  subject=id / type=exch covb corrw;
run;
```

GLMM

```
proc mixed data=data;
      class group time id;
      model y = time group*time / s chisq;
      random INTERCEPT / subject=id;
run;
```

12. Regression: ANCOVA (Equation 7, assumes long data)

Model assumes baseline response, y0, is stored as a separate variable. The data are in long form, however, there is no row for the baseline response. If there is, you will need to exclude from the analysis on the model command line ('if y !=0')

Stata

GEE model

```
xtgee y i.group i.time i.group#i.time y0, fam(gaussian) link(identity)
i(id) corr(exc) eform
```

GLMM

```
mixed y i.group i.time i.group#i.time y0 || id:
```

SAS

GEE model

```
proc genmod data=data;
   class group time id;
   model  y = group time group*time y0  /  dist=normal;
   repeated  subject=id / type=exch covb corrw;
run;
```

GLMM

```
proc mixed data=data;
      class group time id;
      model y = group time group*time y0 / s chisq;
      random INTERCEPT / subject=id;
run;
```