Longitudinal comes from the Latin word longitudo, “length or duration.”

Example of Longitudinal data analysis
What we see in ’Treatment of Lead-Exposed Children Trial’?

- The mean blood lead levels at each measurement occasion for a random subset of 100 children, broken down by treatment group, are presented in Table 2. As expected, due to randomization, the mean response at baseline is similar in the two treatment groups. However, there are discernible differences in the patterns of change in the mean response over time. A graphical presentation of the mean blood lead levels at each occasion is displayed in Figure 1. Note that at week 1 there appears to be a dramatic drop in initial blood lead levels among the children treated with succimer. However, this is followed by a rebound in blood lead levels, as lead stored in the bones and tissues is mobilized and a new equilibrium is achieved. In contrast, for the children treated with a placebo, the trend in the mean response over time is relatively flat.

What we see in ’Six Cities Study of Air Pollution and Health’?

- The girls were measured annually until high school graduation (approximately at age eighteen) or loss to follow-up, and each girl provided a minimum of one and a maximum of 12 observations. At each examination, pulmonary function measurements were obtained from simple spirometry. The basic maneuver in simple spirometry is maximal inspiration followed by forced exhalation as rapidly as possible into a closed chamber. A widely used measure computed from simple spirometry is the volume of air exhaled in the first second of the maneuver, FEY_1 . Although children were measured approximately annually, the data are highly unbalanced when age, rather than chronological time, is used as the metameter for lung function growth. Figure 2 displays a time plot, with joined line segments, of $\log(FEV_1/height)$ versus age for 50 randomly selected girls. Because each girl is not measured at the same age, construction of plots of the mean response versus age can pose difficulties due to sparseness of data at any particular age.

What we see in ’Influence of Menarche on Changes in Body Fat’

- In this data set there are a total of 1049 percent body fat measurements, with an average of 6.4 measurements per subject. The numbers of measurements per subject pre- and post-menarche are approximately equal, with 497 measurements for the pre-menarcheal period (producing an average of 3.1 measurements per subject) and 552 measurements for the post-menarcheal period (producing an average of 3.5 measurements per subject) . In this sample the average age at menarche was 12.8 years . Figure 4 shows a time plot of the individual response profiles (where time is relative to the individual age at menarche). This graph reveals some information about the greater variability of measurement times before menarche. However, it is difficult to discern whether the changes in percent body fat in the pre-menarcheal period are s imilar to the changes in the post-menarcheal period.

What we see in ’Clinical Trial of Anti-Epileptic Drug Progabide’

- The average rates of seizures (per week) at baseline and in the four postrandomization visits are displayed in Figure 5. Note that the mean rates over time in the two treatment groups vary from approximately 3 seizures per week to approximately 5 seizures per week. Thus the expected count of the number of seizures in any 2-week interval is approximately 6 to 10. As we will discuss later, the fact that the mean number of seizures is relatively large has implications for the accuracy of the approximation used in the PQL method.

The above are all examples of longitudinal data. Here is an example of clustered data 'Connecticut Child Surveys'.

- There is now accumulating evidence that the rates of psychiatric disorders in children are substantial, with reported population prevalence rates of childhood psychopathology ranging from 12% to 22%. However, children are considered to be unreliable in reporting on their own psychopathology. As a result many contemporary surveys of childhood psychopathology use proxy informants, usually a child's parent (or primary caregiver) and teacher, to report on the child's psychiatric status. In numerous studies the agreement among multiple informant reports on the child's psychopathology has been found to be poor. It is thought that much of this disagreement is less a result of the unreliability of the informant reports than of true differences in children's behaviors and emotions across different situations and settings, notably in the home and school. A central issue in studies of risk factors for childhood psychopathology is utilization of the information obtained about the child's mental health status from multiple sources or informants.

  Data for our example come from two parallel epidemiological surveys that assessed the mental health and service needs of children, aged 6 to 11, in rural and urban communities in Connecticut. The first survey, the New Haven Child Survey (NHCS), was conducted in 1986 and 1987 in New Haven, Connecticut, a predominantly minority metropolitan center. The second survey, the Eastern Connecticut Child Survey (ECCS) was conducted in 1988 and 1989 and replicated the NHCS in a non-metropolitan planning region covering the eastern third of Connecticut. The two studies used comparable survey procedures. In particular, they used parallel questionnaires designed to be self-administered by the children's parents and teachers. Children's emotional and behavioral problems were assessed with the Child Behavior Checklist (CBCL) and the Teacher's Report Form (TRF), 118-item symptom inventories covering problems commonly seen in child guidance clinics. The CBCL and TRF scales do not provide diagnoses of psychiatric disorders; instead, they provide broad-band measures of emotional (or "internalizing") and behavioral (or "externalizing") disturbance. The CBCL and TRF scale scores can be dichotomized at published clinical cut-points.

  Thus the New Haven Child Survey and the Eastern Connecticut Child Survey provided both a parent's and a teacher's report of psychiatric disturbance in the child as assessed by parallel forms of a standardized psychiatric symptom checklist. These data provide multiple source (here, from two sources: the parent and teacher) information on the psychiatric outcome variable of interest. Of note, these data are cross-sectional but the two sources of information about each child's psychopathology are likely to be positively correlated. Thus, data from the Connecticut Child Surveys are **an example of clustered, but not**

longitudinal, data. In this setting, unlike a typical longitudinal study, the major interest of the analysis is not in changes in the response over time. Instead, the major focus of the analysis is on the effects of subject-specific covariates on the outcome. Figure displays the social and demographic characteristics of the children and the overall rates of externalizing disturbance as determined by CBCL and TRF scale scores in the clinical range.

1 Properties of Longitudinal data analysis

- Responses are not independent, as each subject has repeated measurements.
- In longitudinal design, these repeated measurements are collected over a finite number of visits.
- Thus, we need to consider the dependencies among the observations coming from a given subject.
- This dependence may even depend on some external factors. One needs to understand the data and may even discuss it with other experts before forming the covariance structure.
- There may be missing observations.
- The missingness itself may exhibit interesting patterns, as they may depend on external factors.

The statistical inference relies on the 'correct' formulation of the underlying data-generating mechanism, either probabilistically or in terms of an objective function. Hence, it is important to pay attention to the inherent characteristics of the data.

In most parts, we use R for running different analyses. However, the main focus will be on the underlying assumptions and careful assessment of those assumptions.

Frequency distribution for variables from the Connecticut Child Surveys.

Variables	Count	Percent
Parent informant ($N = 2501$)		
Externalizing		
0 = Normal	2112	84
1 = Borderline/clinical	389	16
Teacher informant ($N = 1428$)		
Externalizing		
0 = Normal	1159	81
1 = Borderline/clinical	269	19
Area		
1 = Rural	874	35
2 = Suburban	428	17
3 = Small city	386	15
4 = Large city	813	33
Single parent		
0 = No	1982	79
1 = Yes	519	21
Child's health		
0 = Good health	1329	53
1 = Fair/bad health	1172	47
Child's gender		
0 = Female	1284	52
1 = Male	1207	48