Chapter1_Draft

"Jessica Yu"

10/21/2021

Introduction

In statistics, there are a plethora of data types that are used to produce different analyses. One in particular is longitudinal data. If we want to track trends and changes over time, such as an effect of a certain drug on the body or growth of a company, longitudinal data and analysis will help us examine those points of interest. Not only can we observe change over time in individuals, but we can look at higher-level grouping, such as change in schools, counties, and organizations. It should be emphasized that only longitudinal data can capture changes within a subject or group; cross-sectional is another type of data in which responses are captured at only one occasion that are compared to other subjects. Ultimately, it cannot provide information about changes over time.

One key aspect of longitudinal data is that there needs to be repeated measurements of the same individuals across multiple periods of time. If there aren't repeated observations, then it is not possible to make any comparisons between two or more time points. Having repeated measurements of the same individual allows for removal of potential confounding effects, such as gender or socioeconomic status, from the analysis.

The measure that captures the observed changes within an individual is referred to as a response trajectory. There are different ways of comparing response trajectories. For example, it is possible to compare the post-treatment vs baseline changes across multiple treatment groups, or it is also possible to compare the rate of change. The method chosen depends on the specific question of the study.

Apart from comparing just the response trajectories, it is also of interest to compare individual differences in the relationship between covariates and the response trajectory. This can be captured using various different statistical models. The choice of model depends on several characteristics of the data.

Characteristics of longitudinal data

While the only requirement of longitudinal data is that there is more than one observation for a given individual, there are other components that affect the models and study design chosen. Data can be unbalanced or balanced: balanced data refers to when all individuals have the same number of repeated measurements taken at the same occasions. In addition, data can also be missing, resulting in automatically unbalanced data. This affects the accuracy of how changes over time are analyzed depending on if there are any patterns to the missing data or not.

Another unique characteristic of longitudinal data is that repeated measurements of each individual are positively correlated. This feature violates conditions of other common statistical methods such as linear regression, where measurements are assumed to be independent. This positive correlation allows for more accurate estimates of the covariates and response trajectories since there is reduced uncertainty knowing that a previous measurement can help predict the next one.

Alongside correlation, covariance between two measurement responses is a crucial measure to calculate. Both measures capture the linear dependence between two measurements depending on covariates, but the correlation is a standardized calculation that does not have units and is simpler to interpret. Typically in

longitudinal analysis, a covariance matrix is calculated for each individual and all of their measurements. The diagonals of this matrix represent the variance of each of the measurements, which are not constant over time. The off-diagonals of the matrix are non-zero to account for the lack of independence between measurements, but are also not constant to account for the assumption that correlations between measurements decrease over time. There are different covariance pattern structures that are imposed that account for these features.

There are several trends that correlation values take on in longitudinal analysis. They are rarely 0, but also rarely 1. They are positive and decrease with longer time separation. These features serve as the underlying premise to the idea that variation can be separated into three distinct parts: 1) between individual variation, 2) within individual variation, and 3) measurement error.

Between individual variation helps explain why measurements from the same individual are more likely to be positively correlated than measurements to a different individual. Within individual variation helps explain why correlations decrease with increasing time differences, and measurement error explains why correlations are never 1. These three types of variation may contribute to total variation in unequal amounts, but may not need to be differentiated depending on the type of longitudinal analysis desired.

→ Notation

Linear models for longitudinal data → more technical Q: how do we decide between them?

As mentioned previously, there are multiple ways to model longitudinal data. When the response variable is continuous and drawn from a multivariate normal distribution, we can consider a model that relates the mean response and the covariates in a linear way. In a linear model all components ~~are~~ can be represented using vectors and matrices. The mean response is modeled by covariates and their slope. Slope values are estimated using maximum likelihood estimation, which maximizes the joint probability of the random variable occurring based on the observed data. Aside from creating line plots to visually capture the mean responses for each individual, there are 3 methods for linear models: 1) response profile analysis, 2) parametric time model, 3) linear mixed effect model.

subsection for each

In response profile analysis, we allow for arbitrary patterns in the mean response over time. The goal of the analysis is to find whether the mean response differs from the interaction of group x time across individuals. An unstructured covariance model is assumed, and requires that the data be balanced. This analysis does not consider time-order of the measurements and does not distinguish between between individual variation and within individual variation. The drawback of response profile analysis is that it can only provide a broad analysis of whether there are differences across groups and time, but does not provide the amount of detail usually needed to answer research questions.

Parametric time models are able to capture time order of the data by fitting linear or quadratic curves to capture an increasing or decreasing pattern over time. Unlike response profile analysis, parametric time models are able to handle unbalanced and missing data. Additionally, there is flexibility in choice of the covariance model; there are several options such as Toeplitz or compound symmetric that impose various structures on the model. However, parametric time models do not difference between subject and within subject variation.

To model the different types of variation that explain correlated measurements among individuals, we can employ the linear mixed effects model. This model distinguishes between fixed effects, which are population characteristics shared by all individuals, and subject specific effects which pertain to each individual. These subject specific effects mean that parameters are random, which induces a structure onto the covariance model. This model is well suited for unbalanced data.

Conclusion for intro? outline rest of thesis?