Regression analysis for longitudinal data

Authors: Vanessa Moulton, Dara O’Neill, Alison Park and George B. Ploubidis

Methods of analysis of data from longitudinal studies allow us to make use of their rich data and to explore the temporal relationships between measures collected across different life stages. Regression analysis is an important and widely-used technique for exploring the relationship between an outcome (e.g. later-life health) and possible explanatory variables (e.g. early-life circumstances). We can gain important insights in social science, biomedical and health research by studying a range of factors throughout the life course, including physical and mental health, and socioeconomic and behavioural factors

In this module you will learn about:

- The advantages of longitudinal data over cross-sectional data analysis
- How to explore a longitudinal dataset and prepare it for analysis
- How to apply general linear, logistic and multinomial regression techniques

Challenge level: advanced

Key concepts:

- Answering research questions with a longitudinal dimension
- Preparing data for longitudinal data analysis
- Examining associations between outcomes and potential explanatory variables
- Adapting analyses for different types of outcome variable
- Updating and comparing statistical models

Suggested citation: Moulton, V., O’Neill, D., Park, A. & Ploubidis, G.B. (2020). *Regression analysis for longitudinal data.* CLOSER Learning Hub, London, UK: CLOSER

Authors: Vanessa Moulton, Dara O’Neill, Alison Park and George B. Ploubidis

1 Introduction and overview

This section introduces some of the important fundamentals of analysing data from longitudinal studies and describes how regression techniques can be used to explore variables relating to different points in an individual’s life course.

1.1 Analysing data from longitudinal studies

The utility of longitudinal studies and the differences between longitudinal and cross-sectional designs are described more fully in the Learning Hub’s Introduction to Longitudinal Studies. There are data analysis methods that allow us to make use of the rich data collected by longitudinal studies and to explore the temporal relationships between measures collected across different life stages. Each of these is suited to the analysis of different types and combinations of variables. Some variables are continuous (e.g. age) and others are categorical (e.g. a list of occupations). We call categorical variables with two levels ‘dichotomous’ (e.g. deceased or living) and, where they are coded as 0 or 1, we can also call them ‘binary’. This guide will teach you about different analytic approaches to exploring how certain types of outcomes are associated with potential explanatory factors.

Dissimilar outcomes can occur even among people who share the same characteristics. The term ‘heterogeneity’ is often used to refer to differences like these. Longitudinal data can help control for such differences by including a wide range of explanatory variables across the life course in statistical models. The problem of ‘omitted variable bias’ is also improved by using longitudinal data, but always remains, as there are connections between the outcome and explanatory variables that have not or could not be included as they are unmeasurable.

Suggested citation: Moulton, V., O’Neill, D., Park, A. & Ploubidis, G.B. (2020). *Regression analysis for longitudinal data*. CLOSER Learning Hub, London, UK: CLOSER

Authors: Vanessa Moulton, Dara O’Neill, Alison Park and George B. Ploubidis

We will use an extract from the National Child Development Study (NCDS) CLOSER Training Dataset to illustrate some of the different methods that can be used in analysing longitudinal data. The NCDS is a cohort study of people born in England, Scotland and Wales during a single week of 1958. In the NCDS, detailed information has been collected on participants from childhood, through adolescence into early adulthood and later life, allowing us to look at different outcomes and potential explanatory variables.

Measurements that have been collected over time include assessments of physical health (e.g. Body Mass Index (BMI) measured at ages 7, 11, 16, 23, 33, 42 and 50), as well as a series of mental health (e.g. Malaise inventory), socio-economic position, and behavioural factors (e.g. smoking), measured at ages 23, 33, 42, and 50. These measures are examples of the variety of data available in the NCDS and other longitudinal studies.

1.2 Overview of this guide

In the following sections, we will present a variety of longitudinal data techniques you can apply to longitudinal data and repeated measures. First, we will explore and prepare the dataset before demonstrating how to apply general linear, logistic and multinomial regression approaches which are commonly used in the analysis of longitudinal study data. In future updates to this module, we will also illustrate how to transfer data to a format suitable for repeated measures analysis. We will also be adding guidance on techniques for analysing such repeated measures data, including multilevel regression, fixed effects, and latent growth models.

We will guide you through these methods as performed in the STATA statistical software package, and we will provide documented syntax to explain the steps involved. Guidance for other statistical software packages is forthcoming.

Suggested citation: Moulton, V., O’Neill, D., Park, A. & Ploubidis, G.B. (2020). *Regression analysis for longitudinal data*. CLOSER Learning Hub, London, UK: CLOSER