

Types of Double Cross-Sectional or Longitudinal Data (Panel Data) and how to Deal with Them

Ameer Kamil Hamzah

Technical Medical Institute, Middle Technical University, Mansour, Iraq

Email id: Ameeralkk@mtu.edu.iq, ameeralkk@yahoo.com

Abstract: *There are many types of data that depend on the type of research study and the type of observations taken for the purpose of the study and also depend on the type of research or research study and the desired results from the analysis of observations. One of these types of data is double longitudinal data or the so-called cross-sectional data. It is a group of ramified chains that go through the same conditions and are studied and analyzed at the same time. The idea of this data is to take observations for a certain period of time and for a certain phenomenon, such as if this phenomenon is a satisfactory condition or currency exchange rates if the subject of the study is economic or, for example, studying a type of agricultural crop that The study was agricultural, etc. The cross-sectional data is characterized by the fact that it takes the case for the studied observations and for a group of variables for the same time period, and requires that these observations have passed through the same conditions that all variables go through and for the specified time period. It generates for us what is scientifically known as cross-sectional or longitudinal double data, which works to address this leakage by conducting a scientific-statistical analysis in accordance with the type of cross-sectional data and study data as explained above, and we will include in this article some types of cross-sectional data and cases that may pass with it.*

Keywords: *panel data, time series, regression, dropout of data*

1. INTRODUCTION

Double longitudinal or cross-sectional data (panel data)

This type of data is considered in one view as both longitudinal and time series data, and panel data is defined as that data whose observations are recorded for (n) cross section over a specified time series. Sections according to the type of phenomenon studied, such as forming certain countries or regions in countries or companies or cases of patients staying in a specific hospital or cases of patients recovering through giving a specific treatment ... etc. Where one phenomenon is studied for all those sections with the same number of approved explanatory variables and for a period of time Equal and here we get the type of balanced longitudinal data, and the researchers' interest in the quality of observations led to the emergence of major types of them , Cross-sectional longitudinal data, including time-series data, and both types give a limited space to determine the nature of the phenomenon and its effects on future prediction, which led researchers to interest in another type of data called double data, especially in the fields of medicine, economics and sociology, where it provides an improved

method for dealing with the phenomenon because researchers obtain indicators Much better than what we obtain by analyzing the time series or the longitudinal section separately. In the case of studying the data at different times or losing one of the data of one of these times, we get the unbalanced double data (unplance panel data), so the cross-sectional data is a set of series temporal^{1,2}

Within certain times and these times determine the work or response of the time series and for the period of study of the phenomenon, the field of research for all cross-sectional data, but when part of this data is lost, the problem of leakage occurs (dropout of data)³.

Objective of the article

Because of the emergence of multiple types of data, the article aims to familiarize researchers with the quality of the studied data, specifically cross-sectional or double data, in order to facilitate the process of tabulating and typesetting them in tables, and then conducting the statistical analysis process.

The problem when studying this type of data and how to deal with it ⁽²⁾

When studying longitudinal cross-sectional data, as it is called (panel data), we sometimes see a case of data leakage, and data leakage usually occurs in three forms shown below.

- 1- A random data loss that affects the response component (the dependent variable)
- 2- Random data loss that does not affect the response element (the dependent variable)
- 3- Loss of response for the dependent variable, so the loss of data is not random

There are two types of longitudinal cross-sectional data, the first, which is, for example, when studying a diseased case for a group of patients and in the same time periods, we obtain data called balanced cross-sectional data (plance panel data). One of the time sections of the treatments of one of these patients shows us the state of leakage in the data, which is called unbalanced cross-sectional data (unplance panel data) and here the problem occurs when studying and the process of leakage is called the studied data.

Theoretical part

There are two types of data leakage depending on the type of relationship and the effect on the response variable (Y_i) (dependent variable)

The model of probability dropout of data (the model in general).

$$Y_i = X_i \beta + Z_i V_i + \epsilon_i \quad \dots (1)$$

Whereas:

- (i) singula $(n_i * 1)$ response vector = Y_i
- . known $(n_i * p)$ destine matrix = X_i
- . unknown $(n_i * 1)$ vector of parameters the social = β
- . known $(p * 1)$ design matrix = Z_i
- . $N(O, \sum v)$ unknown distribution $(r * 1)$ vector of parameters = V_i
- . $N(O, \sum \epsilon)$ independent random residual vector with distribution = ϵ_i

Type of dropout of data

(DROP OUT COMPLETELY AT RANDOM) (DCAR).

Here the dropout of data for (M) from the longitudinal sections is so random that it does not affect the response variable $(Y_i)^2$.

(DROP OUT AT RANDOM) (DAR)

Here the dropout is random for M from the longitudinal sections so that it affects the response variable (Y_i)

For example, the loss of longitudinal cross-sectional data, such as a group of patients lying in the hospital and taking a treatment together, the interruption of one of the patients for a period of time for treatment may affect the degree of response to recovery, and here the importance of the three points mentioned in the research problem appears, and thus the problem of data leakage or losing part of the cross-sectional data appears and thus affects On the process of estimating the parameters as described in the objective of the research.

Note that the regression parameters of the dual data are estimated for the random parameter regression model (RCR) and the general random coefficient regression model (GRCR) by the traditional statistical method, which is the method of least squares LSE (Lest Square Error) and therefore Random feature estimators for each cross-section can be predicted in the presence of the paired data⁴.

It is worth noting that when the marginal tendencies change across the cross-sections of the paired data, these parameters are given the stochastic character, which gives these models the designation of regression models with double data with random parameters and characterized by cross-sectional random errors of heterogeneity, for example, when studying an environmental phenomenon related to air pollution in Iraq and neighboring countries The same problem shows the importance of the study according to cross-sectional data models, and it is not necessary that the tendencies of each country be (a section), as each country has a different atmosphere. At that, we get unrealistic estimates of the phenomenon in Iraq⁴.

Double cross-sectional or longitudinal data regression models

There are two types of regression models, the first is expressed as fixed data in that its effect is fixed, and the second is random effect⁴⁵.

The fixed effect of fixed limits

The constant effect of the parameters pertains exclusively to the fixed limits, without the stochasticity of these two-data models occurring to change the fixed limits through cross-sections over units of time (t)

Whereas:

$$B_1 \ddagger B_2 \ddagger B_3 \ddagger \dots \ddagger B_n$$

whereas

(n) represents the number of cross sections

(BO) represents Fixed border

So the fixed effect model is written as follows.

$$H_0 : B_1 = B_2 = \dots = B_n$$

H1 : at least tow are not equale

The structural homogeneity hypothesis of the fixed effect should be tested using the F. test. If the hypothesis (Ho) is accepted, which states that there is a relationship between the

independent variables and the (dependent) response variable, it means that there is a fixed effect and it becomes one integrated sample. If the hypothesis (Ho) is not accepted, the parameters of the fixed term are estimated according to the dummy variables model, the covariance analysis model, or the least squares model for the variable, which is one of the names of the double data model³.

Second: the random effect of fixed limits.

The random effect of the fixed limits means that the variances of random error are different across the longitudinal sections (or the same thing), so the tests are on the variances of random error, and the equality of variances is tested according to the Hassman test using the chi-square test (X^2).

The hypothesis is as follows

Ho : null hypothesis (fixed effect)

H1 : alternative hypothesis (random effect)

$$X^2 = (bF - bR) \sum^{-1} (bF - bR)$$

When accepting the null hypothesis (Ho): this indicates constant effects, and when rejecting (Ho), it indicates random effects, meaning accepting the alternative hypothesis (H1).

As a general view, the cross-sectional data is seen in one look with the time series data without separating them, but the process of taking, analyzing and studying the graphic series is different.

2. REFERENCES

- [1] al-mufrigi. rihab kathim hamaza , assistant teacher, mohammad sadiq (2018) Estimation of random parameters in paired data regression models with practical application, Doctoral thesis, University of Baghdad / College of Administration and Economics
- [2] rasheed.pr.dr.thafer hussin, ,nazik jaafar, assistant teacher, University of Baghdad / College of Administration and Economics ,(Estimation of the longitudinal data model in the event of information leakage) , Research the publication of the Fourth International Scientific Conference of the Union of Arab Statisticians.
- [3] al-karaghwly,ameer kamil, assistant teacher , abdulrazak, pr.dr.kanaan abdullateef,(Analysis of some nonlinear generalized autoregressive models heterogeneity of the lower order with practical application), 2014 Master's Thesis .
adi H.Baltage , econometric analysis of panel data , published book (2005) .
ohamed Reda Abonazel , (2019), Generalized Estimators Of Stationary Random-Coefficients Panel Data Models: Asymptotic And Small Sample Properties , Cairo University, Egypt .