ANALYSIS OF PANEL DATA

An Introduction

datascience@berkeley

Introduction to Panel Data

What Is Panel Data?

- Panel data, which are often referred to as longitudinal data, have both cross-section and time series dimensions.
- Panel data can be created by sampling the same individuals, families, departments within a company, companies, schools, cities, counties, and so on, over time.
- This gives us both the characteristics and response of interest over multiple time points.
- The example shows two individuals observed daily over a 10-day period.

	Reaction	Days	Subject
1	249.5600	0	308
2	258.7047	1	308
3	250.8006	2	308
4	321.4398	3	308
5	356.8519	4	308
6	414.6901	5	308
7	382.2038	6	308
8	290.1486	7	308
9	430.5853	8	308
10	466.3535	9	308
11	222.7339	0	309
12	205.2658	1	309
13	202.9778	2	309
14	204.7070	3	309
15	207.7161	4	309
16	215.9618	5	309
17	213.6303	6	309
18	217.7272	7	309
19	224.2957	8	309
20	237.3142	9	309

Potentials and Capabilities

- Analysis of panel data provides potentials and capabilities to address questions that would not have been possible using cross-section data.
- Specifically, with multiple observations per subject, we can understand behavior dynamic by observing the same subjects over time.
- We can also understand how these dynamics are related to other variables.
- Within-individual change is characterized in terms of some appropriate summary of the changes in the repeated measurements on each individual during the period of observation.

Characteristics of Panel Data and Implications

Two Key Characteristics:

- A common feature of repeated measurements on an individual is correlation, that is, knowledge of the value of the response on one occasion provides information about the likely value of the response on a future occasion.
- Another common feature of longitudinal data is heterogeneous variability, that is, the variance of the response changes over the duration of the study.

Characteristics of Panel Data and Implications

Consequences:

- These two features of longitudinal data violate the fundamental assumptions of independence and homogeneity of variance that are at the basis of many standard techniques (e.g., *t* test, ANOVA, and multiple linear regression).

Solution:

- To account for these features, statistical models for longitudinal data have two main components: a model for the covariance among repeated measures, coupled with a model for the mean response and its dependence on covariates.
  - Covariance means both the correlations among pairs of repeated measures on an individual and the variability of the responses on different occasions.
  - Failure to properly account for the covariance results in hypothesis tests and CIs that are invalid and may result in misleading inferences.

Berkeley SCHOOL OF INFORMATION