Class 02: Econometrics Data Types

Introduction

In econometrics, the type of data you work with is crucial for determining the appropriate analytical techniques and models. The four main types of data used in econometrics are Cross-Sectional Data, Time Series Data, Pooled Cross Sections, and Panel/Longitudinal Data. Each type of data provides unique insights and requires specific methods for analysis. This lecture will explain each type of data, provide numeric examples, and discuss their applications in econometrics.

1. Cross-Sectional Data

Definition:
Cross-sectional data is data collected at a single point in time across different entities. These entities could be individuals, households, firms, or countries. Cross-sectional data provides a snapshot of various characteristics at a specific moment, allowing for comparison across different entities.

Example

Suppose you conduct a survey on a particular day, collecting data on the age and income of different individuals. Here's how the data might look:

Person	**Age**	**Income ($)**
A	25	50,000
B	30	60,000
C	22	45,000
D	40	70,000
E	35	65,000

Interpretation

This table represents cross-sectional data because it shows the age and income of different individuals at a single point in time. Cross-sectional data is often used to analyze differences between groups or to assess relationships between variables, such as income and education level.

2. Time Series Data

Definition:
Time series data consists of observations of the same variable (or variables) collected at regular intervals over time. This type of data helps track trends, cycles, and other time-dependent patterns, making it essential for forecasting and understanding how variables evolve over time.

Example

Consider a dataset tracking the average monthly temperature in a city over 5 years:

Year	Average Temperature (°C)
2018	15.2
2019	15.5
2020	15.1
2021	16.0
2022	16.2

Interpretation

This table shows time series data because it records the temperature of the city over several years. Time series data is valuable for identifying trends (e.g., a gradual increase in temperature), seasonal patterns, and cyclical movements. In econometrics, time series analysis is used for forecasting economic indicators, analyzing stock prices, and more.

3. Pooled Cross Sections

Definition:
Pooled cross-sectional data combines cross-sectional data from different time periods, with different entities surveyed at each period. This allows for the comparison of different groups across time, helping to identify changes and trends in the data.

Example

Imagine we collected data on housing prices in two different years, 2010 and 2020. The houses are different in each year, but we can still analyze the data to understand how housing prices have changed over time:

Year	House ID	Price ($)	Rooms
2010	1	200,000	3
2010	2	250,000	4
2020	3	300,000	3
2020	4	350,000	4

Interpretation

This table is an example of pooled cross-sectional data because it combines data from two different years, with different houses surveyed in each year. Economists use pooled cross sections to study the effects of policy changes, inflation, or other time-dependent factors by comparing different cross sections over time.

4. Panel/Longitudinal Data

Definition:
Panel data, also known as longitudinal data, tracks the same entities over multiple time periods. This allows for the study of how variables change over time for the same subjects, offering insights into dynamics that cannot be captured with cross-sectional or time series data alone.

Example

Consider a dataset where we follow the same 3 individuals' income over 3 years:

Person	**Year**	**Income ($)**
A	2018	50,000
A	2019	52,000
A	2020	55,000
B	2018	60,000
B	2019	62,000
B	2020	65,000
C	2018	45,000
C	2019	47,000
C	2020	50,000

Interpretation

This table represents panel data because it tracks the income of the same individuals over multiple years. Panel data is powerful for analyzing individual behaviors, detecting causal relationships, and controlling for unobserved heterogeneity. It's commonly used in studies of labor economics, finance, and public health.

. Summary of R Code: Econometrics and Economic Data

A) Cross-Sectional Data

Data File: `wage1.csv`
Purpose: Analyze wage data and other characteristics of individuals at a single point in time.

Key Steps:

i) **Data Loading:** The `wage1.csv` dataset is loaded, which contains cross-sectional data on wages.

ii) **Data Summarization:**

   a) The entire dataset is summarized using the `stargazer` function to generate descriptive statistics.

   b) Specific variables such as `wage`, `lwage` (log of wage), `educ` (education), `exper` (experience), `tenure`, `married`, and `female` are selected and summarized separately.

iii) **Data Selection:** The dataset is modified to retain only the selected variables.

iv) **Frequency Table:** A frequency table is created for the `female` variable to explore the gender distribution.

B) Time Series Data

Data File: `prminwge.csv`
Purpose: Analyze time series data on minimum wages and related economic variables over time.

Key Steps:

i) **Data Loading:** The `prminwge.csv` dataset is loaded, which contains time series data on minimum wages, coverage, unemployment rate, and GNP.

ii) **Data Selection:** Only relevant variables (`year`, `avgmin`, `avgcov`, `prunemp`, `prgnp`) are retained for analysis.

iii) **Data Summarization:**

   a) The dataset is summarized to understand the distribution of the selected variables over time.

   b) A frequency table for the `year` variable is created to check the time coverage of the data.

C) Pooled Cross Sections

Data File: `hprice3.csv`
Purpose: Compare housing prices and characteristics between two different years (1978 and 1981).

Key Steps:

i) **Data Loading:** The `hprice3.csv` dataset is loaded, containing pooled cross-sectional data on housing prices.

ii) **Data Selection:** Key variables (`year`, `y81`, `price`, `lprice`, `rooms`, `baths`) are selected for analysis.

iii) **Data Summarization:**

   a) The dataset is summarized to explore the distribution of housing prices across the two years.

   b) Separate summaries are provided for the years 1978 and 1981 to compare the average prices before and after 1981.

iv) **Frequency Table:** A frequency table is created for the `year` variable to understand the distribution of observations across the two time periods.

D) Panel/Longitudinal Data

Data File: `wagepan.csv`
Purpose: Analyze the panel data of individual wages over multiple years to understand how wages change over time.

Key Steps:

i) **Data Loading:** The `wagepan.csv` dataset is loaded, which contains panel data on individual wages and other variables.

ii) **Data Selection:** The dataset is refined to include variables such as `nr` (individual identifier), `year`, `lwage`, `exper`, `educ`, and `hours`.

iii) **Data Summarization:**

   a) The dataset is summarized to examine the structure and distribution of the variables.

   b) A frequency table for the `year` variable is created to explore the temporal coverage of the panel data.

Overall Code Structure

The code is organized to perform the following operations for each type of data:

- Load the dataset.
- Select and retain relevant variables.
- Summarize the data using descriptive statistics.
- Create frequency tables for key categorical variables.

This structure provides a comprehensive analysis of different econometric data types, demonstrating how to handle, summarize, and interpret data in R.

Conclusion

Understanding the type of data you're working with is crucial in econometrics, as it guides the choice of analytical techniques and models. Cross-sectional data provides a snapshot at a single point in time, time series data reveals trends over time, pooled cross sections compare different groups across time, and panel data tracks the same entities over time. Mastery of these data types and their appropriate use will enhance your econometric analysis and the insights you can draw from it.