Fall 2023

DIME Analytics

Reproducible Research Fundamentals

September 25-29, 2023

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

i2i DIME
TRANSFORM DEVELOPMENT

Fall 2023

DIME Analytics

Data Analysis Hands On

Benjamin Daniels, Roshni Khincha

DIME Analytics

bdaniels@worldbank.org

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

i2i DIME
TRANSFORM DEVELOPMENT

Fall 2023

DIME Analytics

Data Analysis Hands On

During the training, find all materials in our shared DropBox: here

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

i2i DIME
TRANSFORM DEVELOPMENT

Objective of the Course

In this course, you will learn to leverage secondary data to unearth answers to crucial research questions. The analytical phase is the bridge from raw data to discerning insights.

We will predominantly use **Stata** to explore various analytical techniques. We will use the data resulted from the construction exercise.

Municipality Database:

- municipality_database.csv
- Source: Ookla, Open Street Maps and Humanitarian Data Exchange.

State Database:

- state_database.csv
- Source: Ookla, Open Street Maps and Humanitarian Data Exchange.

Data Exercise

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

i2i DIME
TRANSFORM DEVELOPMENT

Exercise

Task 1 Summary statistics: Describing the central tendencies, dispersion, and shape of a dataset's distribution.

Task 2 Boxplots and histograms: Visualizing the distribution of a dataset.

Task 3 Regression analysis: Understanding the relationships between variables.

Task 4 Visual analysis: Crafting informative visualizations.

Task 5 Correlation analysis: Finding relationships between different variables.

Task 1: Summary Statistics

Objective

In this task, we aim to understand the central tendencies, dispersion, and shape of our dataset's distribution using summary statistics.

Steps

Calculate the summary statistics.

Export the summary statistics table to a readable Word file using [putdocx] or [putexcel] commands.

Task 2: Boxplots and Histograms

Objective

Visualize the distribution of different variables in the dataset using boxplots and histograms.

Steps

Use [graph] to create boxplots and histograms in R.

Identify outliers and understand the distribution of your data through boxplots.

Get a sense of the central tendency, variability, and the shape of the distribution of your data through histograms.

Save the most relevant figures using [graph export]

Task 3: Regression tables report

Objective

Understand the relationships between variables through regression analysis.

Steps

1. Selecting Variables: Select dependent (download and upload speeds) and independent variables (e.g., number of schools). Think why this could have endogeneity problems, but, for the purpose of this exercise ignore it.

2. Building the Regression Model: Utilize [regress] to build your regression model using the municipality database.

3. Analyzing the Model: Use the [return list] and [ereturn list] commands to get detailed information on your model.

4. Add clusters: Add clustered standard errors at state level.

5. Save it: Save all your models to Excel using [estout].

Task 4: Visual Analysis using ggplot

Objective

Using graphs can help bring insights and patterns into a clearer focus. They increase the communicative power of your data, translating numbers into visuals that can be more intuitively understood.

Steps

1. **Create scatter plots:** Use [twoway scatter] to analyze the relationship on the previous task. Explore alternative commands, such as [lowess] or [tw lpoly]. See if you can graph multiple visualizations on the same plot.
2. **Create a bar graph :** Find which states suffered more in terms of connectivity from 2020Q1 to Q4 and plot it.

3. **Interpretation:** Do any other graph that could help you to deduce patterns, trends, and insights and save your graphs.

Graphing and visualization tips

Enriching your Visualizations:

- **Layering:** Incorporate different layers
- **Combining:** Utilize [graph combine] to create a matrix of plots.
- **Themes:** Apply different themes to tailor the aesthetics of your plot.
- **Color:** Consider adding color, size, and shape aesthetics to enrich the information shown.

Graph Types and their Utilities:

- **Scatter Plots:** Visualizing relationships between two continuous variables.
- **Bar Charts:** Distribution of a categorical variable.
- **Histograms:** Distribution of a single continuous variable.
- **Box Plots:** Snapshot of the data's central tendency and spread.

See: https://dimewiki.worldbank.org/Stata_Coding_Practices:_Visualization

Task 5: Correlation Analysis

Objective: Understand the relationships between different variables using correlation analysis. Here, we will focus on analyzing the relationship between social infrastructure and average connectivity speed.

Tasks:

1.Filter the necessary data from the municipality database for correlation testing.

2.Perform a correlation test using the [corr] command.

3.Create a correlation matrix using selected variables from the state database.

4.Visualize the correlation matrix using [graph matrix] function to represent correlations graphically.

Fall 2023

DIME Analytics

Tidying Data Hands On

Benjamin Daniels

DIME Analytics

bdaniels@worldbank.org

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

i2i DIME
TRANSFORM DEVELOPMENT