Data Task Instructions

1. Compulsory Section (100 Points)

1.1 Overview

The data consists of partially cleaned temperature and rainfall readings from the Indian Meteorological Department. The datasets in the "Rainfall" and "Temperature" subdirectories consist of daily rainfall and temperature grid point data from 2009-2013. Rainfall is recorded in millimeters and temperature is recorded in degrees Celsius. Each grid point comes from a 1˚(latitude) x 1˚(longitude) grid covering the Indian subcontinent.

When you are done with Sections 1.2 – 1.4 below, please send in the following items:

- ✓ Well-commented code files in the language of your choice (.do files for Stata, .R or .Rmd for R, .py or Jupyter Notebook for Python, etc.)
- ✓ Final dataset produced from Section 1.2
- ✓ Final graphs and table produced from Section 1.3
- ✓ A short document answering the questions raised in Sections 1.2 – 1.4

1.2 Data Cleaning

The central task of this section is to collapse the grid-level dataset into a district-level dataset. The algorithm to match each grid point to a district is as follows:

(1) take a weighted average of daily temperature;

(2) take a weighted average of daily rainfall;

(3) take a weighted sum of daily rainfall

for all grid points within 100 KM of each district's geographic center. The weights are the inverse of the squared distance from the district center.

"district crosswalk small.csv" in the "Geo" subdirectory consists of district centroids for five Indian states: Gujarat, Kerala, Pondicherry, Punjab, and Rajasthan. For the purposes of this task, only use these districts when matching with the grid points. Note: the state and district names in this file correspond with 1961 definitions. They will not always match modern names.

The final product from this section is a district-level daily dataset from 2009-2013 with temperature, rainfall, and total rainfall variables.

Hint: To determine which grid points lie within 100 KM of each district center, you will have to calculate the distance between every grid point and every centroid. The ***geodist*** *or* ***vincenty*** *commands may be useful.*

1.3 Data Exploration

This section uses the district-level daily dataset created in Section 1.2. We are interested in documenting the monsoon season in various districts.

1.3.1 For the Ahmedabad district in Gujarat, create a time series dataset of daily rainfall, averaged over the five years.

1.3.2 Using this time series, create a scatterplot of rainfall by day. On what day does the monsoon season start in Ahmedabad? When does it end? Indicate your answer on the graph. Format your graphs so that they are publication-quality and save them as .pdf files.

1.4 Estimation and Causal Inference

We are interested in using this dataset to estimate the impact of temperature on yearly district-level mortality in India.

1.4.1 Write out a simple econometric model regressing mortality on temperature.

1.4.2 Suppose you estimate the model using OLS and obtain a coefficient of 0.05 with a standard error of 0.02. Interpret this result.

1.4.3 Is your parameter of interest identified in section 1.4.1's model? Write out some potential endogeneity problems.

1.4.4 How could you use a) fixed effects and b) other control variables to remove

potentially confounding factors?

1.4.5 Suppose you want to use variable Z as instrumental variable (IV) to tackle the endogeneity issue. You are supposed to answer the two questions below under this research scheme.

(1) Write out the identifying conditions that Z should meet;

(2) Write out the two-stage least squares estimation method.

2. Bonus Section (20 Point, Not a Mandatory Exercise to Submit)

Raw Chinese datasets are usually in untidy formats. Suppose that we have a big dataset about Chinese enterprises. "Addresses.xlsx" consists of 50 simplified firm addresses from it, and "Admin 2019.xlsx" includes the names and IDs of all provinces, prefectures and districts in China. Note that these administrative divisions (行政区划) correspond with 2019 definition. Please use this version throughout.

Your task is to use the two datasets to find the administrative divisions of the 50 firms. The final output should be a dataset with the following variables: address, province name, city name, district name, province ID, city ID and district ID. As much you can, write code that works as the data size expands, and so please do not resort to manual labor. For example, do not manually type in "海淀区" as the district name if you see the firm is located in "北京市海淀区".

When you finish this section, please send in the following items:

- ✓ Well-commented code files in the language of your choice
- ✓ Final dataset with the following variables: address, province name, city name, district name, province ID, city ID and district ID