PREDOC Data Task

July 2025

This data task will ask you to summarize, visualize, and analyze pollution data. It will also ask you to think critically about the interpretation of your results. Section 1 asks you to familiarize yourself with the data, merge two datasets, and produce summary statistics describing the data. Section 2 asks you to visualize specific patterns or trends in the data. Section 3 asks you to perform regressions and interpret those regressions. Section 4 explores threats to this interpretation. **Please do not spend more than 5 hours on the task**. Good luck!

Data Background

You will be provided with two files named `ca-ozone-2024.Rdata` and `ca-pm25-2024.Rdata` (or `.dta`). The file contains daily readings from pollution monitors in counties throughout the state of California. All data—save for the the death variable—are retrieved directly from federal air monitoring sites. Refer to the data dictionary at the end of the instructions for a description of the variables.

Section 1: Data Processing and Summarizing

1. For both datasets, the unit of observation is a county-monitor-day group. Merge the two data sets to create a single file. Pre-process the data as necessary to facilitate the merge.

2. Produce a table with the mean, median, minimum, maximum, and standard deviation for ozone, PM 2.5, and AQI for the entire sample.

3. Produce a table with the same statistics for just ozone, but split the sample by the source variable (AQS vs. AirNow). In other words, your table should report the mean, min, etc., for the portion of the data that comes for AQS and those same numbers for the data coming from AirNow. Also include the number of observations for each group.

4. The federal government does not use AirNow data for rule-making:

   > The AirNow data are not fully verified and validated through the quality assurance procedures monitoring organizations use to officially submit and certify data on the EPA AQS (Air Quality System) and, therefore, cannot be used to formulate or support regulation, guidance or any other Agency decision or position.

   Does the table you produced provide any evidence that the data quality may be systematically different? What other evidence would help support this conclusion (discuss, but do not produce, the evidence)?

5. Create a county-day dataset to proceed with the analysis. All subsequent questions will use this new dataset. The data set should contain the three pollution variables, mortality, the county name and code, the date, and the CBSA name and code (one county is always assigned the same CBSA). How did you reconcile different pollution readings for different monitors within the county?

6. How many counties are missing days?

Section 2: Visualizing the Data

1. Produce a plot showing the distributions of ozone and PM 2.5. The distributions should be separate lines, sets of dots, bars, etc, but on the same set of axis. Choose an appropriate type of graph to complete this task and make your graph easy to digest.

   Hint: Pay attention to the scale of the variables. Can we preserve the distribution while standardizing the scales?

2. Produce a time series plot of ozone for Los Angeles county (Code: 037) in the month of February. Do you suspect autocorrelation? [1] How might you test for it (you do not need to test it)?

Section 3: Pollution and Mortality

We will now investigate the relationship between pollution and mortality. The Air Quality Index (AQI) is an index designed to measure the aggregate effect of different pollutants on air quality. We will use that in this section to represent pollution. Note: the mortality data are not real.

1. Estimate the association between pollution and mortality by running the following regression:

$$\text{Mortality}_{it} = \beta_0 + \beta_1 \text{AQI}_{it} + \alpha_i + \alpha_t \tag{1}$$

   where α_i and α_t are fixed effects for county i and time t. Report and interpret β_1. Why do we include fixed effects?

2. I am concerned that yesterday's pollution affects mortality too. Include one lag of AQI and rerun the regression. Report and interpret the coefficients on AQI and yesterday's AQI.

3. I am curious if the relationship differs by whether a county is in a CBSA. Rerun the original regression from Section 3, Question 1, but add an interaction between an indicator for whether a county is in a CBSA and AQI (i..e, $AQI_{it} \times \mathbb{1}_{i \in CBSA}$). Interpret the coefficients.

4. I did not ask you to include a separate variable for whether a county is in a CBSA but just the interaction. If you were to include it, it would be omitted as collinear. What is it collinear with?

Section 4: Discussion on Threats

The following questions ask you to think about threats to identification we have ignored until now. In 8 or fewer sentences, briefly discuss the questions. Do not report additional empirical results or cite external research.

1. Suppose I know pollution follows an AR(3) process and thus the partial autocorrelation is greater than zero for the three previous days. In other words, today's pollution has a non-zero and positive relationship with pollution yesterday, the day before, and the day before that, even if I control for all other day's pollution when identifying that relationship. How many lags of pollution should I include in the regression to avoid omitted variables bias? Does it matter how these lags associate with my outcome?

2. We are seeking to answer the question: Does exposure to pollution increase mortality? Suppose we had found that there was no effect of lags of pollution and we estimated the coefficient in Section 3, Question 1. Can the estimated result be interpreted causally? Why or why not?

[1] Autocorrelation is when something is correlated with itself over time. Here, it would be the equivalent of saying, "today's pollution is correlated with yesterday's pollution (and possible the day before, or the day before that...)."

3. Suppose instead that pollution's main effect is that it kills people who were going to die tomorrow. This is a mechanism called harvesting.

   Further suppose I identified the set of controls necessary to ensure that the omitted variable bias for all my regressions is zero. If I ran a regression of tomorrow's mortality on today's pollution with those controls, how would the coefficient compare to a regression where I run today's mortality on today's pollution with the analogous set of controls? What about if I took the sum of mortality for today and tomorrow and regressed that on pollution today with the proper set of controls?

1 Data Dictionary

Table 1: ca-ozone-2024

Variable	Definition
date	Date of measurement
siteid	Unique ID of Pollution Monitor
ozone_value	Ozone reading
aqi	Air Quality Index
mortality	Number of deaths (synthetic data)
ozone_source	Type of Data
cbsacode	Core-Based Statistical Area Code
cbsaname	CBSA Name
county_code	County FIPS code
county_name	County name

Table 2: ca-pm25-2024

Variable	Definition
date	Date of measurement
siteid	Unique ID of Pollution Monitor
pm25_value	PM 2.5 reading
pm25_source	Type of Data
county_code	County FIPS code