2022-04-01 Progress on Second Year Paper

Aaron Watt

Using PurpleAir to Replace Missing Pollution Data

- Focusing mostly on faculty feedback from presentation

1. Develop the story of incentives more in the introduction

> Jim: we would like to know more if possible about who has the discretion to decide what data to omit, and what their incentives are... pertains both to explaining the importance of attainment status and who the key stakeholders are to the broader audience, and offering additional insight on the details of who the “data managers” are, who hires them, and whether they are plausibly subject to pressure.

- waiting on a call-back from EPA folks to answer some questions and possibly schedule a short interview/discussion.
- will add more details to introduction

2. How much room is left in raw data for manipulation?

> Another exercise would be to ask how much room is left for additional manipulation. That is, suppose a data controller cared about boosting numbers and wanted to drop as many data points as would be allowed. Given the data that are being reported and the slack in the number dropped, how much of a difference would additional manipulation make? This is sort of a scale exercise for asking how useful manipulation might be.

Reporting Standard: 75% of days in each year, 75% (18) of hours in each day

Easy, ball-park way to implement in raw data: - drop all days down to the lowest 18 hours (75%) - order remaining days from highest to lowest PM - drop the highest days until I reach the 75% minimum reporting requirement

Hard to know if this is the absolute minimum attainable value from dropping observations but seems like it would be very close and probably similar to a heuristic used by someone on the ground.

3. Develop motivation: how manipulatable is this measure?

> Ethan: The possibility of flipping attainment status is one way to turn this into an economics/policy narrative, but you might also consider something more broad, like stating that “effective regulation requires reliable measurement,” and then asking if the measure here is manipulable. Obviously we know that there are other types of manipulation (from prior literature), so your contribution is to look for the possibility of manipulation in this particular mechanism.

- I like this take. Switching "flipping attainment status" language to "how manipulable / reliable is this measure given the amount of legal / allowed omission of data?"

4. Kernel regression?

> Ethan: Isn't a kernel regression the right approach here? Asymptotic performance of current approach (inverse distance) may not work?

- Have discussed a little bit and have more office hours scheduled
- The issue is that PurpleAir sensors are coming in and out so I don't have a balanced panel of PurleAir sensors for each EPA monitor

5. Regression to the mean in prediction regression

> Ethan: First table seems to involve regression to the mean (Purple Air a rhs variable, with measurement error).

- revisiting with Ethan, not sure what is to be done about this.

6. Multiple hypothesis testing

> Ethan: Multiple testing (however many sites...)? If drawing inference that a *particular* site is engaged in manipulation, current version of inference seems problematic.

- Unclear to me what the asyptotic distribution of the design values are (especially the 24-hour DV).
- Ethan and Max both directed me to talk to Michael -- seeing him on Monday.

7. Confidence intervals for non-compliant site highly asymmetric—not clear why?

- Confidence intervalus for Daily DVs are (nearly) symmetric and 24-hour DVs are highly asymmetric.
- This is expected for symmetric prediction intervals used to fill in gaps in real data.
- I added a discussion of this in the appendix and briefly reference it in the result section.