1. **Prompt**

Your task is to use data on the early spread of the coronavirus in 2020 in NYC to identify the channels through which the coronavirus spreads and make some policy recommendations. As a guide, refer to the paper by Almagro and Orane-Hutchinson (2020), who explore the reasons why some neighborhoods are much harder hit by the pandemic than others. Using the 2018 American Community Survey (ACS) and COVID-19 test data from March and April 2020 from the Department of Health and Mental Hygiene of New York City (DOH), they find a relationship between different occupations and the spread of the virus.

Disparities in the incidence of testing for COVID-19 and in access to tests have been documented by Borjas (2020) and Schmitt-Grohé et. al (2020). Selection into testing which varies by neighborhood may confound the true rate of positives. You may wish to consider several different dependent variables as measures of infection rate. Almagro and Orane-Hutchinson (2020) use the fraction of positive tests to date as their measure of infection rate, but you might use the total number of positive tests per capita (which you can construct from the data provided), or deaths per capita (available to download from the DOH starting in June 2020). Could these alternative measures of infection rates provide a more compelling link between neighborhood characteristics and the spread of COVID-19?

In addition, Almagro and Orane-Hutchinson (2020) suggest that the ACS data is more likely to reflect neighborhood characteristics during the early weeks of the pandemic due to the significant economic impact of the city lockdown. How might the difference in the timing of the ACS data and these measures of exposure affect your estimates?

2. **Data**

For the purposes of replication, data on neighborhood characteristics at the zip code level is provided from the 2018 American Community Survey (ACS). This is combined with COVID test data from the Department of Health and Mental Hygiene of New York City (DOH). You may use the data provided or supplement it with publicly available data from the DOH, ACS or other sources to answer the prompt. The data provided in the .dta file include:

- Daily data on number of positive tests, total number of tests and fraction of positive tests (to date) ranging from April 1 to May 14, with April 2 and April 6 missing.
- Neighborhood characteristics from the ACS, such as occupations, race, age, population density, commuting patterns and income. The occupations data are further categorized into 13 groups detailed in Almagro and Orane-Hutchinson (2020).
- The variables 'time', 'lat' and 'lng' are provided in the dataset. The authors use spatial HAC standard errors to allow for spatial autocorrelation in the error terms for each neighborhood. If you wish to compute OLS with spatial HAC standard errors as in Almagro and Orane-Hutchinson (2020), you can use the .ado or .m files provided by Hsiang (2010). Go to http://www.fight-entropy.com/2010/06/standard-error-adjustment-ols-for.html (for the STATA help file), and for the .ado and .m files: http://www.globalpolicy.science/code/ (cite Hsiang 2010 if you use this code). The authors use a distance cutoff of 2km. You may want to consider why might this be more appropriate than a simple heteroskedasticity correction.

3. Guidelines

As stated earlier, the prompt is to **identify a policy relevant causal relationship describing the spread of COVID-19 in NYC .**

The goal of the assessment is to assess your ability to analyze data and identify a causal estimate that is policy relevant, interesting, and well-defended. The goal is not to produce a statistically significant result, nor to produce informative, but causally unidentified descriptives. We want to understand how you think about economics problems, and how you can use the economic and econometric toolkit to analyze those problems. The information provided in the prompt was relevant to COVID-19 in 2021 and should serve as a helpful starting guide, but you should not feel limited to only thinking about what was discussed in the prompt.

The response should contain three portions

1. A description of the policy relevant causal relationship you hope to uncover. This section is where any relevant literature should be cited. Although we do not expect a full literature review, you may find it helpful to review methodologically or topically similar work
2. A description of the data and method you use. You may find it helpful to divide this into two sections: a data, and a method section.
    - **Data:** If you use any additional data sources, describe them here and discuss their value add. Discuss the sample you are using and the sample population of interest.
    - **Method:** Building upon your discussion of the data, describe the method you use to identify the causal effect. We're interested in not just the method you are using, but why you chose it and why it is well suited to your empirical situation. We also want you to think critically about the potential flaws with the methodology that you are using. No method is perfect, and we expect you to understand that. If there is a different method that you would like to use but given the time constraints associated with obtaining that data, you are unable to do so, you are also welcome to discuss that here. We would like to see some evidence that you are able to work empirically with causal inference tools but knowing how to obtain a valid result is more important than obtaining results of poor quality.
3. Results and Discussion. Do the results you have obtained make sense? What might be driving them? Given the time limits, a single table and/or chart may be sufficient. We certainly do not expect robustness checks, although we would value discussion of checks you would do in a hypothetical world.
4. Please state at the end of the paper how much time you spent on the task.

Note on Plagiarism

Please note that the use of ChatGPT or generative AI tool is expressly prohibited. In addition, if we identify or suspect that a team plagiarized a paper from the literature, we will contact the team and that team may be barred from participating in this and future iterations of the econometrics game. We understand that the literature can serve as an inspiration, and that no

idea is truly unique, but there is a difference between being inspired by, and copying the methods section of a paper.

References

Almagro, M., Orane-Hutchinson, A, 2020. JUE Insight: "The determinants of the differential exposure to COVID-19 in New York city and their evolution over time", Journal of Urban Economics, 103293, ISSN 0094-1190.
Borjas, G.J. , 2020. Demographic determinants of testing incidence and covid-19 infections in new york city neighbourhoods. COVID Econ. Vetted Real-Time Pap. (3) 12–39 .
Hsiang, S.M., 2010. Temperatures and cyclones strongly associated with economic production in the caribbean and central america. Proceed. Natl. Acad. Sci. 107 (35), 15367–15372.
Schmitt-Grohé, S. , Teoh, K. , Uribe, M. , 2020. Covid-19: testing inequality in new york city. Covid Econ. Vetted Real-Time Pap. (27) 27–43.