**PPGA503** **HW5 prompt**
2023 Due Nov 26

This assignment builds on HW4, specifically by utilizing more advances techniques to increase robustness and generate more nuanced insights. Please use dataset hw5. Please add the new commands for hw5 onto the .do file you wrote for hw4 (meaning you are extending your existing .do file). You will turn in both a write-up of your analysis and the complete .do file you used for hw4 and hw5.

As a reminder, the dataset is a modified version of the World Bank's **Indonesia Database for Policy and Economic Research** (INDO-DAPOER). You can access the background material in the following places: here and here. The dataset and a list of variables are available on Canvas (under week 11 module). The dataset includes a significant number of variables covering health, education, governance, economic, development, and natural resource attributes at the *district level* in Indonesia. Note that Indonesia, aside from having the fourth largest population in the world (spread across thousands of islands stretching approximately 5k kilometers from west to east), is also among the most decentralized countries in the world. This creates immense variation in both governance and developmental outcomes across the districts.[1] There are slightly over 500 districts in one of two types: *kota* (city) are more urbanized, while *kabupaten* (regency) are typically more rural. Provinces, of which there are just under 40, form the meso (middle) tier of government.

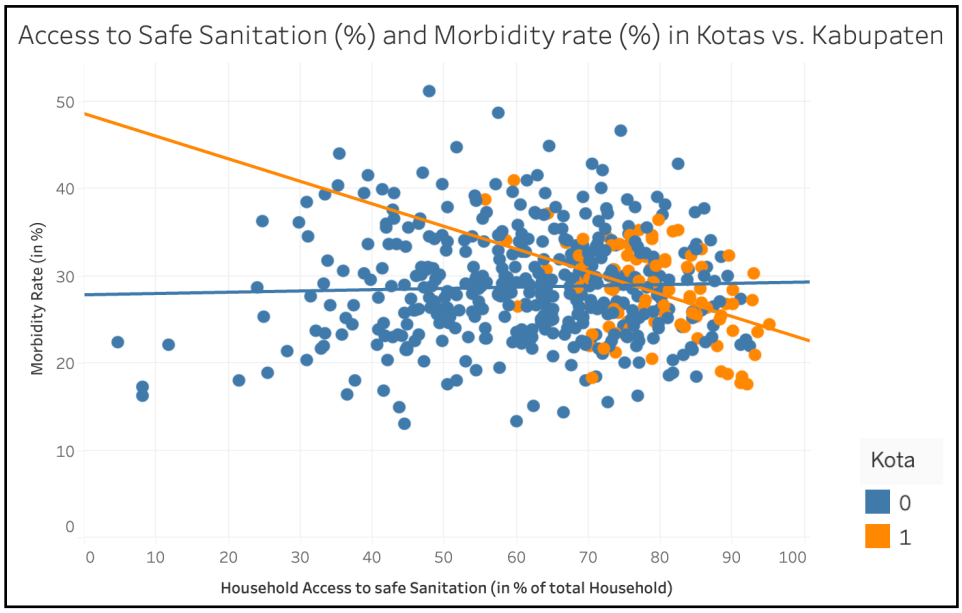
As this is an extension of hw4, we continue to focus on variation in health levels, meaning Morbidity rate (in %) remains the DV. Let's use **model 3** from HW4 as the basis of our analysis. That means we are starting with a base model that contains the following six IVs: (1) proportion of population over 65; (2) poverty rate; (3) physician density; (4) puskesmas density; (5) population density; (6) road density.

1. Let begin by examining how Household Access to Safe Sanitation (%) affects Morbidity rates. In this case, it is reasonable to theorize that access to safe sanitation and its impact on morbidity rates depends on urbanization — see graph below for an initial exploration of this relationship. So, we decide to use the Kota dummy (which takes a value of 1 for kota and 0 for kabupaten) to create an interaction term Kota * Household Access to Sanitation. Make sure you include the right variables!

   Estimate a model (this becomes **model 1** for HW#5) that adds the interaction effect onto our base model (i.e., model 3 from HW#4, see note above)**.** After you estimate this model, use the post-estimation margins command to estimate the two slopes: **margins VAR, dydx(VAR)**. Next, please interpret what you are seeing in a few sentences.

2. As you'll recall, we've noted that OLS can be highly subject to bias from outliers that exert a disproportionate influence on outcomes. Please check for this (use the lvr2plot) to help you. If you are concerned about any observations, take a look at what they have in common. Can you control for what makes them unusual? Please take action if necessary (including dropping severe outliers). This becomes **model 2** in your table. In your writeup, discuss in less that 200 words total: (1) what you found, (2) what you did in response, and (3) how it changed your general conclusions (if it did).

[1] If you are really intrigued by the case and interested in further reading, you could give this a try: "Indonesia's Decentralization Experiment: Motivations, Successes, and Unintended Consequences" by Ostwald, Tajima, and Samphantharak (2016).




3. I'd like you to critically think about our model. Given your (whether intuitive or expert) understanding of variation in health outcomes, is there anything important missing from the model? Go back to the variable list to see whether it can be refined further. You may add up to two additional variables, as you see fit. Estimate the new model and include the findings as **model 3**. Interpret the results in 200 words or less. Be sure to address what you have added (and why), and what impact (if any) it has on your understanding of the determinants of morbidity.

4. Now let's look for violations of the GM assumptions that may be biasing our findings. Assess for heteroskedasticity and multicollinearity, and take whatever corrective measures you think are appropriate. As before, describe in 200 words or less: (1) what you found (and how), (2) what you did in response, and (3) how the corrections affect your model (if at all). If the corrective actions significantly affected your model, please include the new findings as **model 4.**

5. Finally, let's think like policy makers. Let's assume for a moment that a given amount of funding could achieve one of two things:

    a. 10% increase in puskesmas density
    b. 5% increase in physician density

    Using your most recent model as the basis of your analysis, which intervention do you predict would have the greatest effect on decreasing morbidity? (Hint: use the post-estimation **margins command** to specify values. Remember that if you don't specify a value for a given variable, Stata will use the mean value). Write your response in 200 words or less.

6. [Optional for bonus] Is there anything else you'd like to do to improve our understanding of morbidity? Refine your model? Additional analysis or estimates? If so, feel free to do it! If you have new results, you can report them as **model 5**. Give us a short (200-word max) summary of what you did and what you found.

Good luck and hope you find some of this enjoyable (or at least interesting!) ☺