Initial bla bla on the high quality of their work.

Specific topics to consider/suggestions

We didn’t understand why you removed the observations with probability variables (e.g. prbarr and prbconv) having values beyond [0, 1]. These variables are not true probabilities and in theory can be above 1 – for example, an offense can involve more than one offender; if they are all arrested, #arrests/#offense in this case will be above 1. By doing so, you deleted nearly 10 observations, which is a lot in the universe of ~90 total counties in the dataset. And this can also hurt your random sampling assumption.

Recall that the log transformation variables in a regression allows the interpretation of the estimated coefficient as the impact in % change for small % changes.

Figure 3 requires a more detailed explanation. I’m not sure about what each square means.

Wages are typically higher in places that have higher density not because living costs are higher in regions of higher density. Beyond the minimum wage, wages are higher for workers that add more value, and they are typically engaged in activities typically concentrated in regions of higher density – federal workers maybe a valid exception.

About CLM assumptions - Random sampling: as you said, there are 100 counties in NC and your data has only 81. The original dataset was already lacking 9-10 counties, which is in itself a source of concern. In your data exploration you deleted another 9-10. Some discussion on the implications of it should be done… are the deleted counties related in some ways, so their exclusion is biasing the estimates? I guess you can’t tell for those out of the original dataset, but you can for the ones you deleted.

About the sentence: “The key coefficient of -0.059 for prbarr suggests that increasing the probability of arrest by 10 percentage points would be associated with about 5.9 fewer crimes per 1,000 people, holding other factors constant.” - It should be “10%” instead of “10 percentage points” (for the math on the log of prbarr work, you should multiply prbarr by 1.1, not at 0.1 to it).

We missed some further discussion on the meaning of the positive coefficient on log(polpc). As you mentioned at some point, this seems to be related to reverse causality. That being so, should this variable be kept in the model?

We really liked the simple and efficient way you summarized the discussion on omitted variables.

All that being said, very nice job!!