Does high urbanization necessarily associate with low natality in US?

Natality, which denotes the proportion of births relative to the population size, is a crucial metric in demographic research. In this study, we examined the association between urbanization levels and natality across counties in the U.S. We employed both generalized linear models (GLM) and generalized estimating equations (GEE), incorporating various link function families. To conclude, we'll utilize the principal component analysis to identify key covariates that account for the majority of the model's variance.

We collected annual birth data at the county level from 2011-2020 from the Centers for Disease Control and Prevention. Subsequently, demographic data for the counties, detailing populations by gender, race, origin, and age group from 1990-2020, was sourced from the National Cancer Institute. To quantify each county's urbanization degree, we employed the Rural-Urban Continuum Codes (RUCC) provided by the U.S. Department of Agriculture. The RUCC system assigns each county an ordinal value from 1 (representing counties in metro areas with a population of 1 million or more) to 9 (indicating counties that are completely rural or have an urban population of less than 2,500 and are not adjacent to a metro area). A higher RUCC value corresponds to a lower degree of urbanization. Given the relatively stable nature of urbanization levels over a decade, we found it appropriate to use the RUCC data from 2013 as a representative measure.

Upon acquiring the datasets, we undertook several transformations on specific covariates. For instance, we stabilized the variance in the demographic data by taking the square root of each value, followed by logging the population. Additionally, for the year, population, and RUCC data, we adjusted each value by subtracting the mean ($x \longrightarrow x - \text{mean}(x)$), ensuring the data was approximately centered around 0. Subsequently, we amalgamated the three datasets using the county FIPS codes as a unifying key.

The relationship between mean and variance is pivotal, as it informs the selection of appropriate mean and variance structures. To analyze this, we logged the mean and variance of the newborn population within each county and applied an ordinary least squares regression to the transformed values. If the variance is proportional to the expectation, then the log-variance equates to the log-expectation plus a constant, implying a regression line slope of 1. However, our regression yielded a slope of 1.88 with an R-squared value of 0.779. This indicates that the structure $\text{Var}[Y|X=x] = \Phi E[Y|X=x]$ is unlikely to be valid, necessitating exploration of alternative structures.

We subsequently applied the Generalized Linear Model (GLM) with a quasi-Poisson link function to the data. In this model, the response variable is the number of births, while the explanatory variables include the logarithm of populations and the RUCC. The sample encompasses 5,618 data points. Notably, the quasi-Poisson regression extends from the generalized Poisson regression, and it inherently diverges from the assumption that expectation equals variance due to its scaled distribution. As per the GLM regression summary, the RUCC's coefficient is 0.025 with a 95% confidence interval ranging from 0.017 to 0.033. This indicates a significant positive correlation between the RUCC and natality, suggesting that the effect is unlikely to be null. It's important to highlight, however, that the covariance structure in this GLM

model is nonrobust, meaning it doesn't discount any data when calculating covariances between features.

Our datasets encompass numerous data points that exhibit statistical dependencies, which are vital to consider in our analysis. For instance, the number of births in a given year might be influenced by figures from previous years, indicative of serial dependence. Furthermore, natality in one county could be interrelated with that in neighboring counties within the same state. To account for these dependencies, we employed Generalized Estimating Equations (GEE), which extend the GLM framework to accommodate correlated data.

Before fitting the GEE model, it's imperative to select an appropriate link function. We employed both the Gamma and Poisson families to model the relationship between the number of births and RUCC, using the county FIPS code as a grouping variable and the logarithmic population as an offset. Regarding the covariance structure, we opted for the exchangeable working correlation structure. This assumes that any two observations within the same cluster correlate at a level denoted by ρ, a parameter estimated from the data. We visualized the findings by plotting the absolute Pearson residual against the log predicted mean (as shown in Figure 1) for both families. Notably, the slope associated with the Gamma family is closer to 0, suggesting that the Gamma family more accurately reflects the underlying mean/variance relationship.

Despite its utility, our GEE model has limitations, one of which is its simplicity due to the inclusion of only one or two covariates. This restricts its ability to effectively capture and explain the variance within the data. To address this shortcoming, we can implement principal component analysis (PCA). By performing singular value decomposition and identifying the top-ranked principal components, our revised model can be formulated as:

$$\mathrm{Births} \sim (\mathrm{population}+\mathrm{Rucc}) * \mathrm{year} + \sum_{i=1}^{N} p_i$$

The covariate p_i is the covariate which can explain i-th most variance and we'll set N to 10, 20 and 50. We also chose the county FIPS code to be the group and the logarithmic population to be the offset. If N=10, the coefficient of RUCC is 0.0394 and the 95% confidence interval is (0.024,0.054). If N=20, the coefficient of RUCC is 0.0117 and the 95% confidence interval is (-6.86e-05,0.023). If N=50, the coefficient of RUCC is 0.0181 and the 95% confidence interval is (0.007,0.029).

In conclusion, across all our models, the coefficients of RUCC are consistently positive, indicating a direct positive correlation with natality. These coefficients are statistically significant, suggesting that they are unlikely to be null or negative. This implies that as urbanization increases, natality decreases. Such a trend may be attributed to factors like heightened work pressures and the escalating costs of child-rearing in metropolitan areas. Thus, in the context of our study, higher urbanization is linked to lower natality rates in the US.

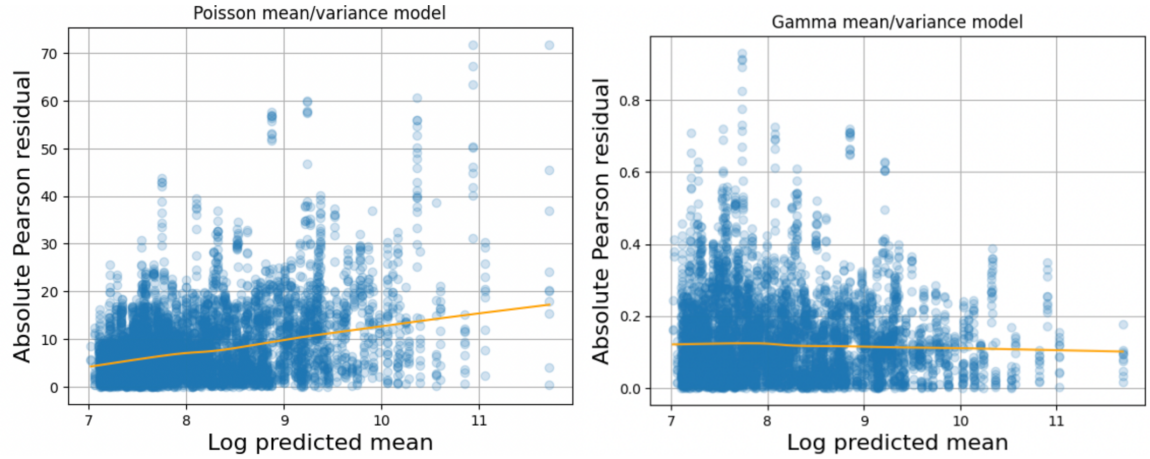

Figure 1: The scatter plot of absolute Pearson residual versus log predicted mean of Poisson mean/variance model and Gamma mean/variance model