Demographic predictors of natality change from 2007-2014 to 2015-2022

Introduction

Different Population demographic structures might have different levels of births. In this report, we try to understand the major influential demographic combination for birth rates in counties in the United States and detect the change of the major influential demographic combination from 2007-2014 to 2015-2022.

Data

We gather 2007-2022 natality from CDC (https://wonder.cdc.gov/natality-current.html), 2016 demographic data including age, gender, race, origin from NIH (https://seer.cancer.gov/popdata/download.html), rural-urban codes from USDA (https://seer.cancer.gov/popdata/download.html). Eventually, rows of the final data represent each county's demographic in 2016, birth counts, and rural-urban codes for every single year. The demographic columns have been taken square roots to stabilize the variance.

Methodology (PCR+GEE)

(1) PCR (Principal Component Regression)

Our dataset contains 304 demographic variables, which are high dimensional and highly correlated. PCR reducing the dimensionality of a dataset by projecting original variables onto a lower-dimensional subspace, will be suitable for our dataset and we only select the principal components to fit the regression.

(2) GEE(Generalized Estimating Equations)

To specify the appropriate GLM, we explored the mean/variance relationship. Since natality is a count and is expected to have a multiplicative mean structure and positive mean/variance relationship, we limited our consideration to GEE models of the form below:

$$\log(\mathrm{E}[\mathrm{Y}|\ \mathrm{X}=\mathrm{x}]) \ = \ \log(\mathrm{pop}) + \beta_0 + \beta_{\mathrm{rucc}}\mathrm{RUCC} + \beta_{\mathrm{year}}\mathrm{Year} + \beta_1\mathrm{x1}\ldots + \beta_{\mathrm{p}}\mathrm{x}_{\mathrm{p}}$$

$$\mathrm{Var}[Y|X=x] = \phi \cdot E[Y|X=x]^p$$

$$R_i(\alpha) = \alpha \cdot 1_{n_i \times n_i} + (1-\alpha) \cdot I_{n_i \times n_i}.$$

In the mean equation, Population should be set as offset, since the birth rate makes more sense. RUCC, Year represents the rural-urban codes and year respectively. x1…xp represents the p principal component we selected.

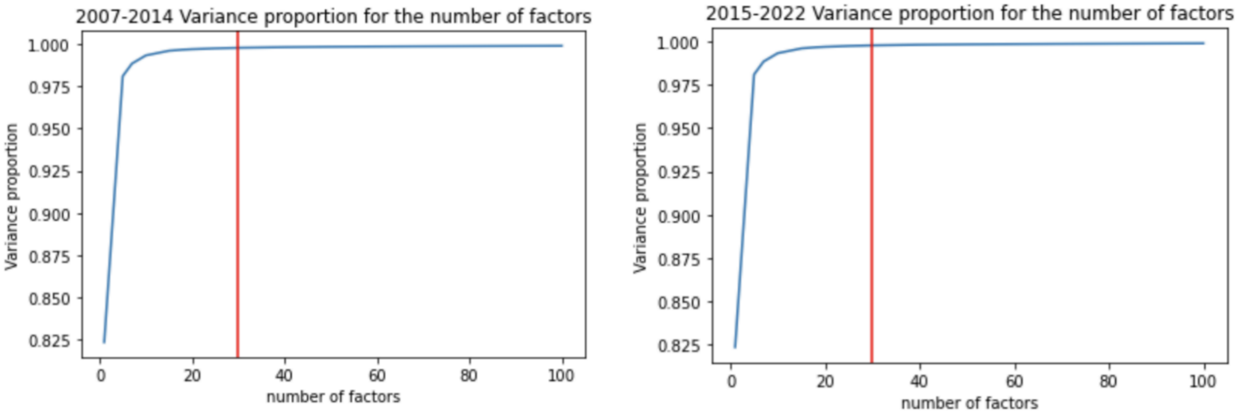
Result

(1) Goodness of fit

Graph 1 shows the explained variance proportion as the number of factors increase. We could conclude for both groups of 2007-2014 and 2015-2022, the appropriate number of factors should be no less than 30. That's because, after 30 factors, the proportion of explained variance is close to 1 and also becomes stable.

Graph 2 shows Gamma variance diagnostics with 30 factors. The proper value of p in the variance structure should be 2. Gamma variance diagnostics show approximate horizontal lines for 2007-2014 and 2015-2022, telling us our variance-mean structure is appropriate. Also, with 30 factors being chosen, the Gamma variance diagnostics would be closer to constant.

Graph 1 Cumulative percentage of explained variance



Graph 2 Gamma variance diagnostics

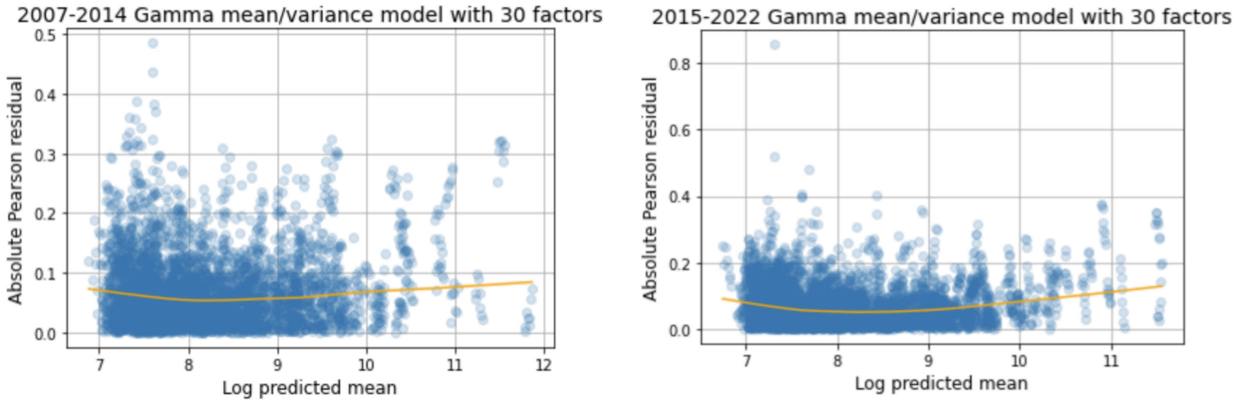


Table 1 illustrates the score test comparing models with different numbers of factors for 2007-2014 and 2015-2022. The list shows 50 factors should be a more proper number of factors, because p-values are always less than 0.05 until comparing 60 factors versus 50 factors. However, 30 factors would also be fine, given the Gamma variance diagnostics is more appropriate.

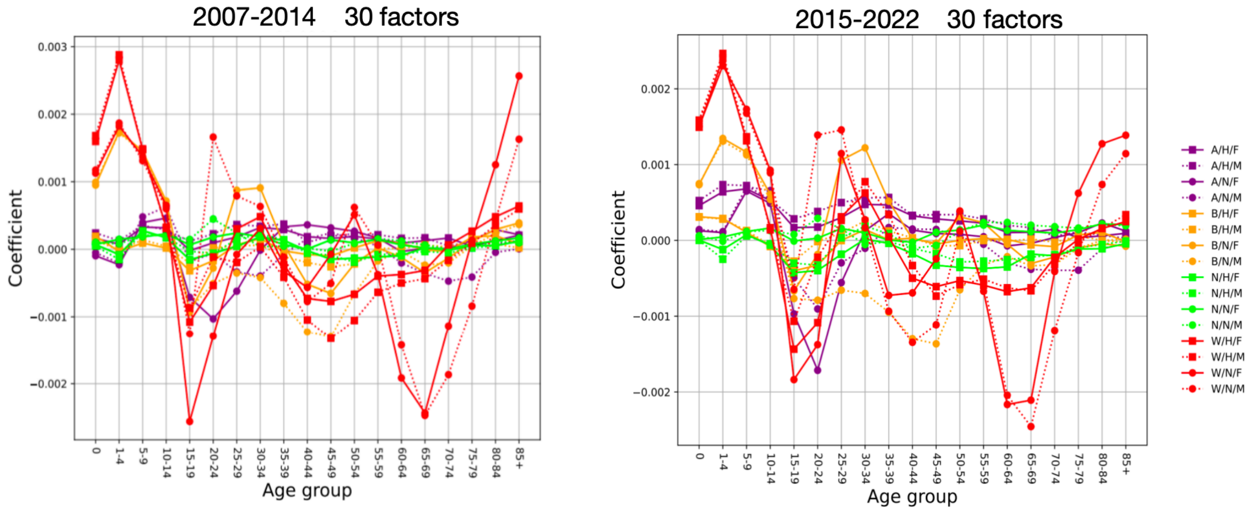
Table 1 Score Test

	2007-2014	2015-2022
10 versus 1	0.0000	0.0000
20 versus 10	0.0000	0.0000
30 versus 20	0.0403	0.0403
40 versus 30	0.0068	0.0068

50 versus 40	0.0030	0.0030
60 versus 50	0.3617	0.3617

(2) Comparison of major influential demographic combinations for two periods.

Graph 3 Coefficient of Demographic Combinations



For the general patterns of different demographics, from 2007-2014 to 2015-2022, the absolute value of A/N/F coefficient magnitude increased dramatically for multiple age groups compared to others. The coefficient absolute values of all the Native races(green line) also increased a little for nearly all age groups. In contrast, the demographic with White race(red line) coefficient magnitude decreased for some age groups with large magnitude during 2007-2014. The black race's coefficient magnitude didn't change a lot in these two periods. This indicates that Asians and Native play more and more important roles in birth rates, while White race's influence wanes even though they still dominate for nearly all age groups.

Focusing on the specific age group, for the 15-19 group, W/N/F coefficient, which was the highest during 2007-2014 decreased its magnitude by more than 0.005. In the 20-24 group, we could find A/N/F coefficient absolute value increases by more than 0.005.

Another interesting pattern is that, for the 1-4 age group, during 2007-2014, the Asian's coefficients are generally negative, while during 2015-2022, the Asian's coefficients become generally positive. This means during 2007-2014 more Asian in 1-4 age people, less birth rate in that county. However, during 2015-2022 more Asian in 1-4 age people, more birth rate in that county.