Investigating County-Level Natality Trends: Temporal and Demographic Insights

Vincent Weng

Introduction

This memo investigates how birth counts vary across U.S. counties from 2011 to 2020 and examines the extent to which these variations can be explained by county-level attributes. This includes demographic composition, rural-urban classifications, and socioeconomic indicators. Using various statistical methods, this analysis rigorously accounts for heteroscedasticity and temporal correlation within counties to ensure robust conclusions. By integrating natality data from the CDC with demographic and socioeconomic datasets, this study aims to identify key drivers of natality trends and assess disparities across counties. Understanding the factors that influence natality patterns across counties over time is crucial for informing public health policies and resource allocation.

Methods

We began by filtering for counties with natality records and reshaping the demographic data to align population counts by race, ethnicity, sex, and age. A log transformation was applied to stabilize variance, followed by centering to remove mean effects across counties and groups. These steps reduced noise, addressed heteroscedasticity, and prepared the data for dimensionality reduction via singular value decomposition. To assess demographic complexity, we examined the decay of singular values, which revealed approximately 10–12 dominant components captured most of the variation, with the remainder following an exponential or power-law decline. Biplots of the leading components visualized the main demographic gradients across counties, clarifying patterns by age, sex, and ethnicity while highlighting county-level outliers.

For modeling natality trends, we merged the birth data with demographic, socioeconomic, and geographic covariates. We used log-transformed population as an offset to account for the expected scaling between population size and birth counts. We also assessed mean-variance relationships to justify modeling overdispersion. Then, we fit generalized estimating equations (GEE) with a Gamma family to account for heteroscedasticity and within-county correlation over time. Predictors included urbanicity (RUCC), deprivation (ADI), time trends, and interaction terms. To further account for demographic structure, we applied principal components regression (PCR). After extracting orthogonal components via SVD from the centered demographic data, we incrementally added these components to the regression model. This allowed us to capture key patterns while avoiding multicollinearity. Finally, score tests were used to evaluate how many components significantly improved model fit.

Results

We first explored the demographic structure of U.S. counties using SVD on age-race-sex population data. The singular value spectrum (Figure 1) showed a steep initial drop followed by a slower decay, indicating that approximately 10–12 components capture the majority of demographic variation. This multiphasic decay pattern is consistent with both exponential and power-law behavior, which supports the use of low-rank approximations for dimensionality reduction. Biplots of the leading components (Figure 2) revealed major demographic gradients, with clear separations by age, race/ethnicity, and sex. Next, we assessed the relationship between population size and birth counts. As shown in Figure 3 (left), birth counts scale nearly one-to-one with population size on the log scale, justifying the use of log population as an offset in regression models. However, variance increased more rapidly than the mean, with a slope of roughly 1.88 in the log-log mean-variance plot (Figure 3, right), indicating substantial overdispersion. This motivated the use of a Gamma variance model rather than a standard Poisson model.

To evaluate the roles of urbanicity and socioeconomic status, we modeled birth counts using GEE with predictors ADI, RUCC, and time. Poisson regressions suggested both ADI and RUCC were significant predictors, but RUCC lost significance when accounting for within-county clustering. Switching to a Gamma variance structure yielded improved residual behavior (Figure 4), and model diagnostics confirmed a better fit. Adding a linear time trend revealed a consistent decline in natality from 2011 to 2020, particularly in high-deprivation counties. Final model estimates with interaction terms (Figure 5) confirmed that time-varying effects were statistically meaningful.

Finally, we applied PCR to quantify the influence of demographic structure on natality. By incrementally introducing principal components derived from SVD, we examined how model behavior changed with increasing demographic complexity. Figure 6 shows that with 5 components, the estimated effects across age-race-sex groups were smooth and interpretable. However, adding more components—such as 30 and 60—introduced more detail as well as greater variability, particularly in smaller subgroups. Score tests (Figure 7) provided formal model comparison results. They showed statistically significant gains in model fit up to 60 components. Beyond that, the improvement plateaued, with the comparison between 60 and 70 components yielding a non-significant p-value ($p = 0.41$). This suggests that models with more than 60 components may begin to overfit the data. Overall, PCR with approximately 60 demographic factors strikes a balance between explanatory power and model parsimony.

Interpretation

This analysis reveals that demographic structure plays a central role in shaping county-level natality trends. The sharp decay in singular values suggests that most demographic variation can be captured with a relatively low number of components, indicating strong underlying structure across age, race/ethnicity, and sex. The PCR results further support this, showing that models with around 60 components strike a balance between detail and stability, capturing nuanced effects of specific subgroups. This was particularly evident for younger populations and certain racial/ethnic groups on birth counts. These patterns reflect persistent demographic gradients in natality that are likely due to broader social, cultural, and economic differences across counties.

In addition to demographic factors, we found that socioeconomic disadvantage, as measured by ADI, is consistently associated with lower natality rates. This relationship strengthened over time, with more deprived counties experiencing sharp declines in birth counts between 2011 and 2020. Urbanicity, while initially significant, had weaker effects after accounting for within-county clustering. This suggests that deprivation may be a more proximate driver. Together, these findings highlight both the stability of demographic effects and the growing influence of structural inequality on birth outcomes.

Overall, this study underscores the value of combining dimension reduction with robust longitudinal modeling to identify the complex drivers of natality. The use of principal components enabled us to capture demographic variation without overfitting, while generalized estimating equations addressed non-constant variance and temporal dependence. All in all, these methods provide a scalable framework for monitoring natality patterns and understanding how shifting population structures and socioeconomic conditions shape public health trajectories at the county level.

Figures

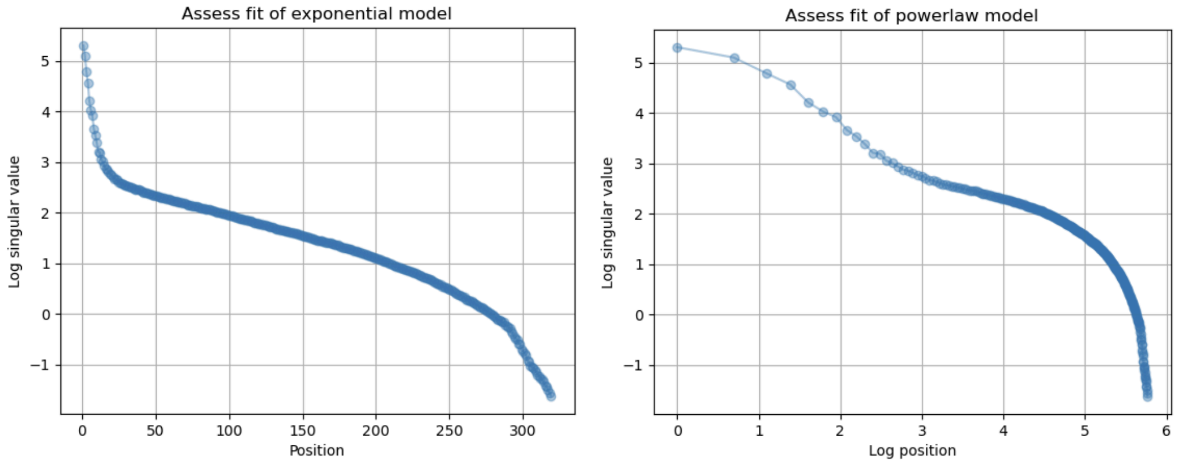
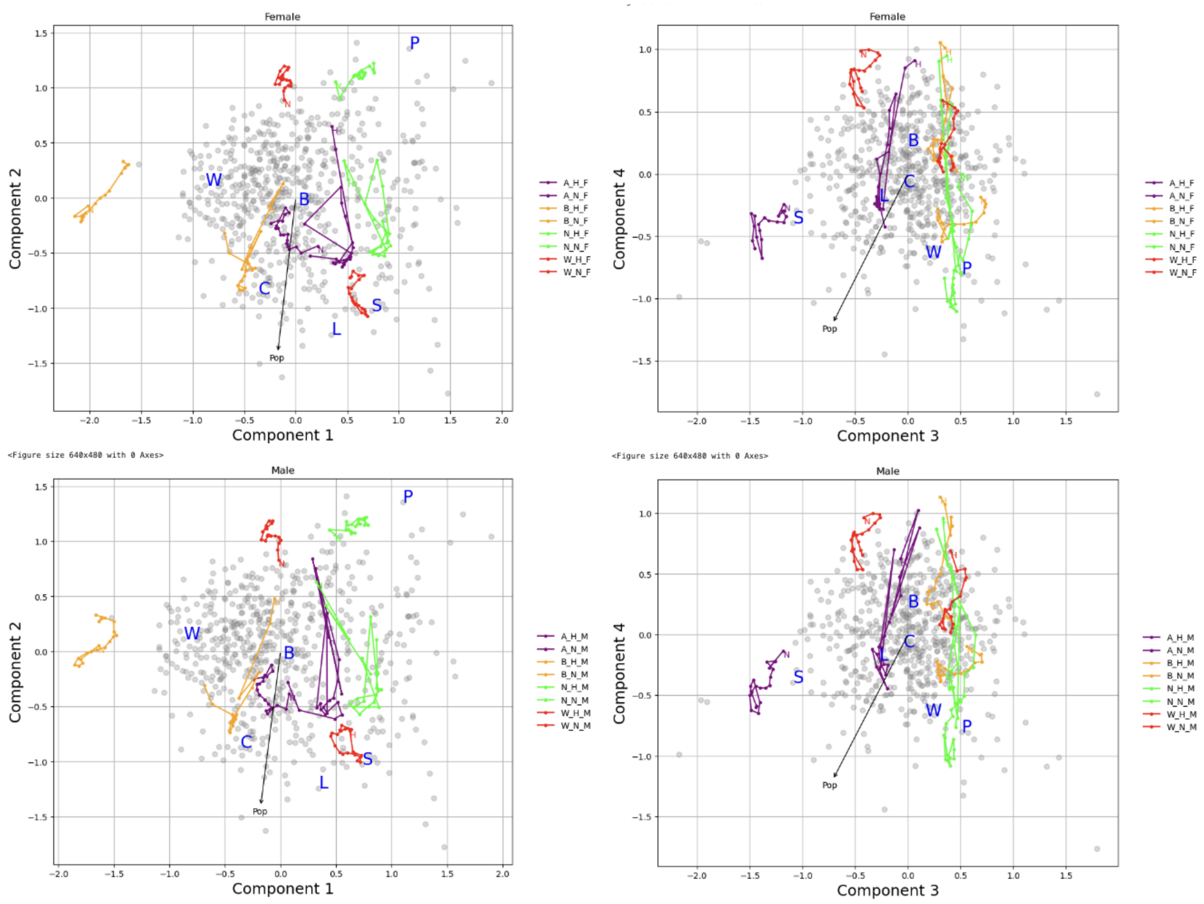


Figure 1: Decay of Singular Values: Exponential vs. Power-Law Fit



```
<Figure size 640x480 with 0 Axes>
```

```
<Figure size 640x480 with 0 Axes>
```

Figure 2: SVD Biplots of County Demographics by Sex

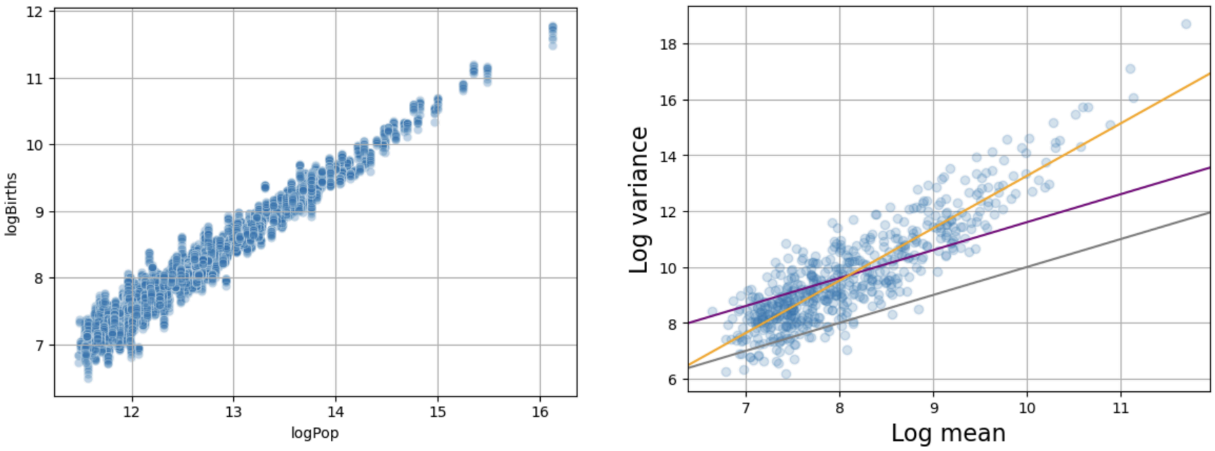


Figure 3: Relationship between log population and log births (left), and log mean vs. log variance of birth counts (right), supporting offset use and motivating a Gamma model.

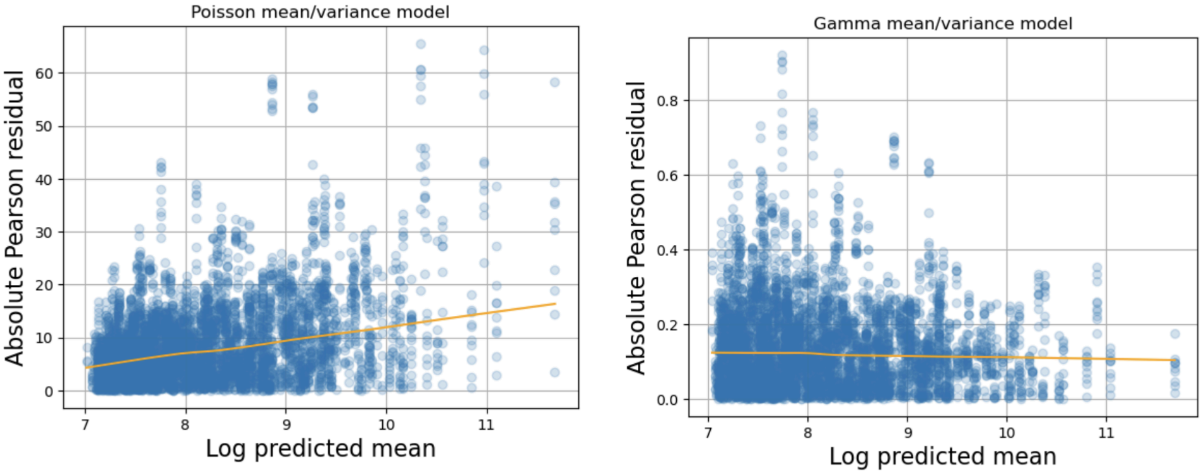
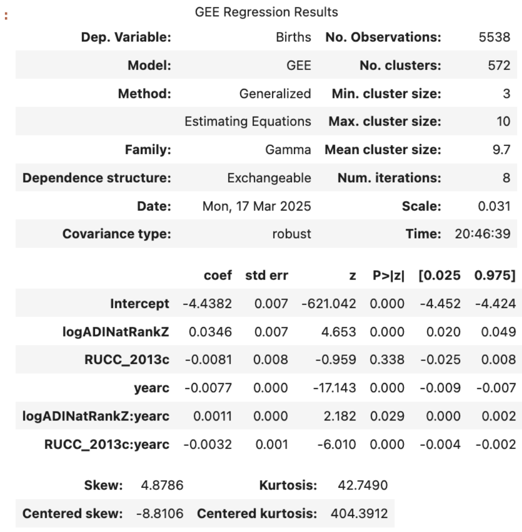


Figure 4: Diagnostic plots comparing Poisson and Gamma mean–variance relationships



```
                          GEE Regression Results
        Dep. Variable:                 Births   No. Observations:      5538
                Model:                    GEE        No. clusters:      572
               Method:            Generalized   Min. cluster size:        3
                           Estimating Equations   Max. cluster size:       10
               Family:                  Gamma   Mean cluster size:      9.7
 Dependence structure:           Exchangeable     Num. iterations:        8
                 Date:       Mon, 17 Mar 2025              Scale:     0.031
      Covariance type:                 robust               Time:  20:46:39

                          coef  std err        z  P>|z|  [0.025  0.975]
            Intercept  -4.4382    0.007 -621.042  0.000  -4.452  -4.424
      logADINatRankZ    0.0346    0.007    4.653  0.000   0.020   0.049
          RUCC_2013c   -0.0081    0.008   -0.959  0.338  -0.025   0.008
               yearc   -0.0077    0.000  -17.143  0.000  -0.009  -0.007
logADINatRankZ:yearc    0.0011    0.000    2.182  0.029   0.000   0.002
    RUCC_2013c:yearc   -0.0032    0.001   -6.010  0.000  -0.004  -0.002

           Skew:   4.8786          Kurtosis:   42.7490
  Centered skew:  -8.8106  Centered kurtosis:  404.3912
```

Figure 5: GEE regression results with Gamma variance and interaction terms

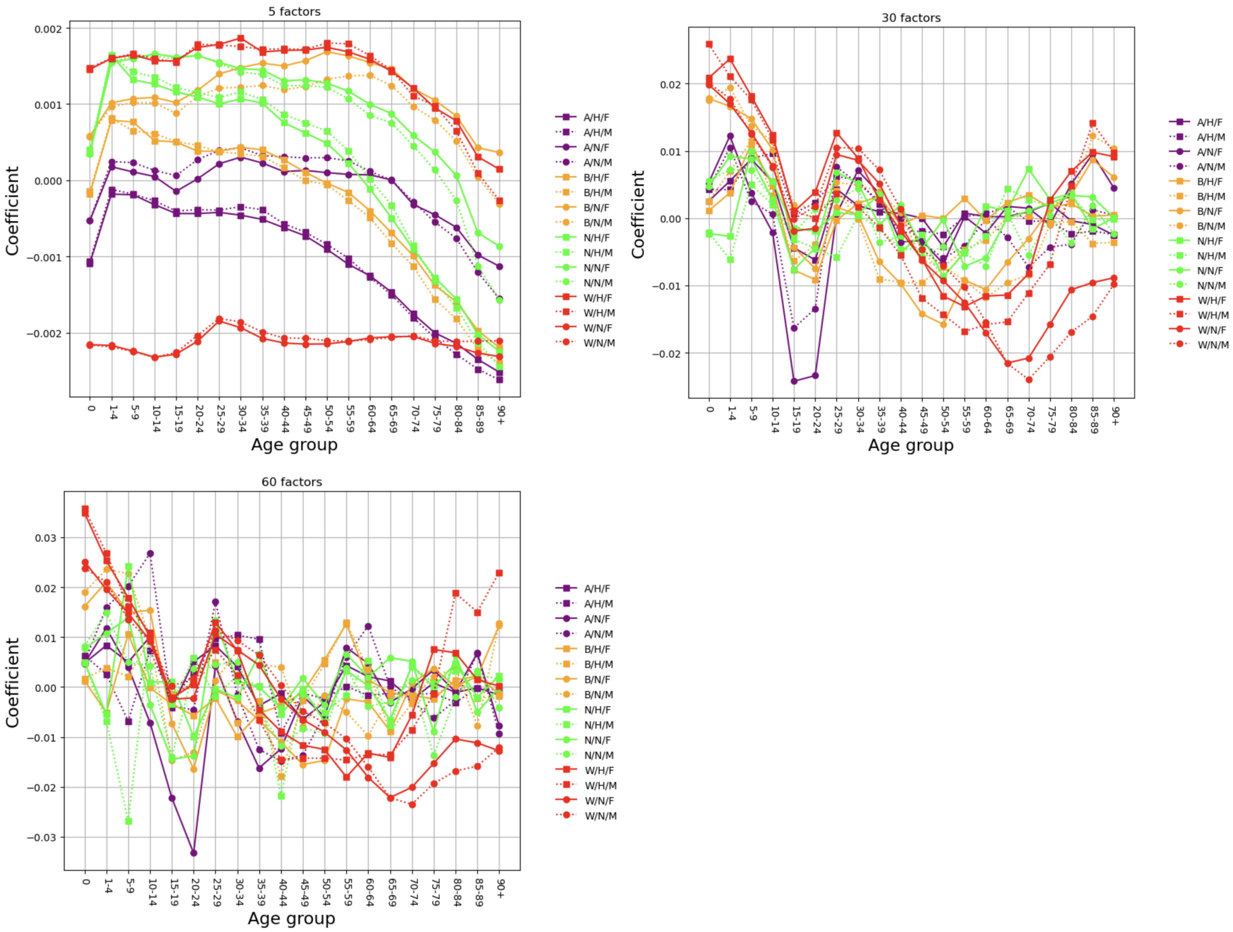
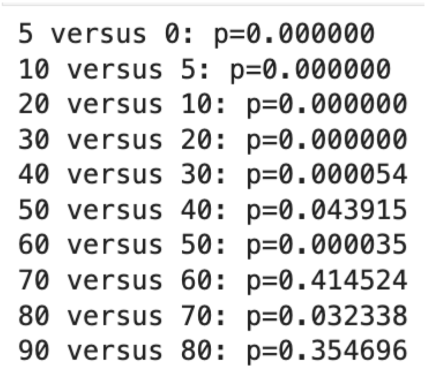


Figure 6: Demographic coefficients from PCR with 5, 30, and 60 factors.



```
5 versus 0: p=0.000000
10 versus 5: p=0.000000
20 versus 10: p=0.000000
30 versus 20: p=0.000000
40 versus 30: p=0.000054
50 versus 40: p=0.043915
60 versus 50: p=0.000035
70 versus 60: p=0.414524
80 versus 70: p=0.032338
90 versus 80: p=0.354696
```

Figure 7: Score tests comparing PCR models with increasing factors.