Youngjun Woo
UMID: yjwoo
February 107th, 2022

Association between demographic variables and birth counts by counties

We aim to analyze the association between demographic variables and birth counts by county using Natality data from the United States. Natality data consists of birth counts information from 2016 to 2020 for a total 626 counties. In addition, it has population information of subgroups divided by 4 demographic variables. The demographic information is as follows.

Race	W	White
	B	Black
	N	American Indian / Alaska Native
	A	Asian or Pacific Islander
Origin	N	Non-Hispanic
	H	Hispanic
Sex	F	Female
	M	Male

Since there are 4, 2, 2, and 19 categories for each demographic variable, a total of 4 x 2 x 2 x 19 = 304 groups exist. That is, we have the population information of these 304 groups for each county. In addition to demographic variables, there is also a numerical variable called RUCC_2013. RUCC_2013 is a variable that shows how urban the county is on a scale of 1 ~ 9. 48 counties have missing values for both population and RUCC_2013 variables. Since we cannot impute these missing values, only a total of 580 counties excluding these counties were analyzed.

Since our problem is the single-index regression and there is some mean-variance relationship on our data, we used GLM. Since we measured birth counts for each county over 5 years (2016 ~ 2020), there is a serial dependency in our data. To accommodate this we can use GEE, which is a framework that extends the GLM approach to accommodate dependent data. As in GLM, we select an exponential family based on the structure of the conditional mean and conditional variance of the dependent variable. To this end, we checked the relationship between the log(sample mean) and log(sample variance) of the birth count data within each county. As shown in Figure 1, we can check that there is a linear structure with Log(sample mean) and Log (sample variance) of the birth count. Also, in Table 1, the coefficient of

Log(sample mean) is almost 2 and the t- statistic is very large (46), so there is statistical evidence that this coefficient is not 0. Since $log(Sample\ variance) = const + 2log(Sample\ meean)$, the original relationship is $Sample\ variance = Sample\ mean^2$ Therefore, we set gamma distribution as our exponential family for GEE based on this mean-variance relationship.

OLS Regression Results

Dep. Variable:	var	R-squared:	0.773
Model:	OLS	Adj. R-squared:	0.773
Method:	Least Squares	F-statistic:	2129.
Date:	Fri, 10 Feb 2023	Prob (F-statistic):	2.75e-203
Time:	17:31:01	Log-Likelihood:	-878.73
No. Observations:	626	AIC:	1761.
Df Residuals:	624	BIC:	1770.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-6.5135	0.349	-18.675	0.000	-7.198	-5.829
mean	1.9543	0.042	46.141	0.000	1.871	2.037

Omnibus:	50.967	Durbin-Watson:	1.714
Prob(Omnibus):	0.000	Jarque-Bera (JB):	72.723
Skew:	-0.617	Prob(JB):	1.62e-16
Kurtosis:	4.125	Cond. No.	73.9

Table1: OLS regression results of log(Mean) ~ log(Var) of the birth count variable to check the mean and variance structure of our data

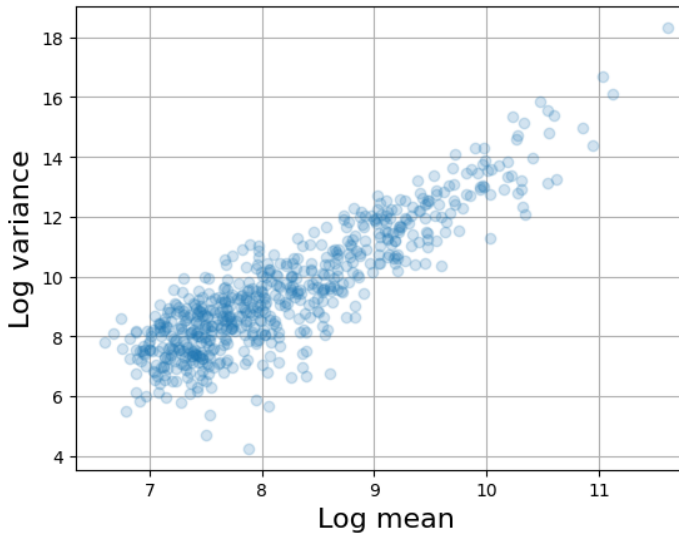

Figure1: Scatter plot of log(Mean) and log(Var) to check the mean and variance structure of our data

In addition, since the populations of 304 demographic subgroups are used as a covariate, the dimension of our covariate is high. To handle the challenge of high-dimensional regression, we used the PCA method. In Figure 2, it can be seen that the first PC variable explains about 80% of the total variance, and each subsequent PC variable explains less than 10% of the total variance. That is, it can be confirmed that the information from the total 304 variables can be successfully represented with fewer PC variables almost without losing information.

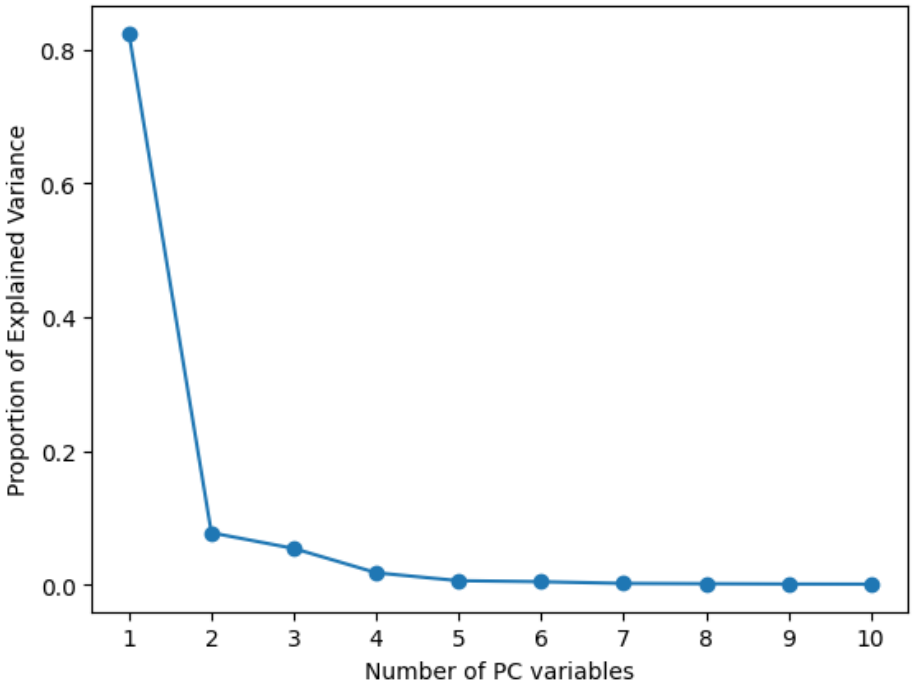

Figure 2: PVE plot of PC variables

How many PC variables to choose was determined by comparing models that used different numbers of PC variables. Table 2 below shows the score test results between different models. As a result of the score test, after including 50 PC variables, there is no statistical difference between the models even when additional PC variables are included.

H0	H1	p-value
Model with 0 PC variables	Model with 10 PC variables	P = 0.000
Model with 10 PC variables	Model with 20 PC variables	P = 0.000
Model with 20 PC variables	Model with 30 PC variables	P = 0.036
Model with 30 PC variables	Model with 40 PC variables	P = 0.002
Model with 40 PC variables	Model with 50 PC variables	P = 0.003
Model with 50 PC variables	Model with 60 PC variables	P = 0.291
Model with 60 PC variables	Model with 70 PC variables	P = 0.458
Model with 70 PC variables	Model with 80 PC variables	P = 0.654

Model with 80 PC variables	Model with 90 PC variables	P = 0.176
Model with 90 PC variables	Model with 100 PC variables	P = 0.618

Table 2: Result of score test comparing nested models with PC variables

To analyze the association between demographic subgroup populations and birth counts, we converted the coefficients of PC variables into coefficients of original demographic variables and then visualized them by age group. Figure 3 shows the result. The color represents race, the shape represents the origin, and the dotted or solid line represents sex. The x-axis represents the age group. Overall, all subgroups have very small coefficients, indicating that no particular subgroup has a dominant effect on the birth count.

Also, we can check an approximate common trend among several graphs. Many graphs have high coefficients in the group of 5-9 age or younger, and then the coefficients gradually decrease. This can be interpreted as an increase in the population of young children, which means that the possibility of having a new child increases, and thus has an effect on increasing the birth count. Then, the age group between 15 - 19 and 25 - 29 has a high coefficient again. Since people between the ages of 15 - 19 and 25 - 29 usually have children, it can be interpreted that as the population in these age groups increases, the birth count increases. From the 55 - 59 to 75 -79 age group, many subgroups have negative coefficients. Since people of this old age are less likely to give birth to children, it can be interpreted that as the population of this age increases, the birth count decreases.

Also, comparing the coefficients between the races, it can be seen that the white race group (red lines) have a larger absolute coefficient than the other race groups. That is, when the population of the white group increases, the birth count increases or decreases more than when the population of other race groups increases. Conversely, American Indian / Alaska Native (green lines) have generally smaller absolute coefficients than other races. In other words, the increase in the population of this race group has the least effect on the birth count compared to other race groups. Next, comparing the coefficients between origins, the hispanic group in the circle has a larger absolute coefficient than the non-hispanic groups. That is, when the population of the hispanic group increases, the birth count increases or decreases more than when the population of the non-hispanic group increases.

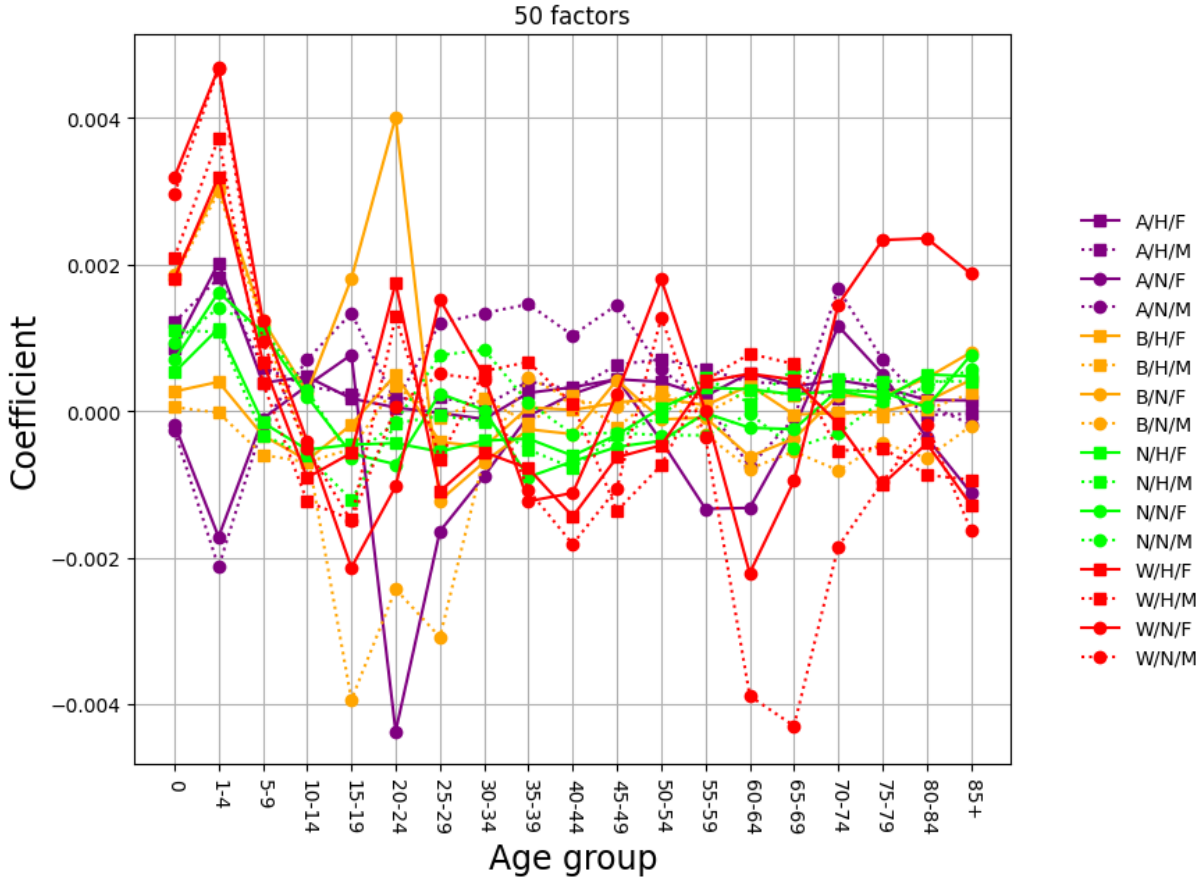

Figure 3: Coefficients of different demographic subgroups' population