Premature births in California

final project, ucla extension

Ekaterina Ivanova

winter2021

## Introduction

Every year premature birth affects about 380000 babies in the United States.

It accounts for more than 70% of neonatal deaths.

It causes long-term neural disabilities in almost 50% of the cases.

Also, preterm birth has a significant impact on quality of life, emotional stress on the family, and healthcare costs.

Preterm births are all live births less than 37 weeks of gestation.

Very preterm births are all live births less than 32 weeks of gestation.

The known causes explain only about one-third of preterm births.

We poorly understand why it happens. There can be genetic polymorphisms and epigenetic factors, infections, stress, vascular disorder. Race, economic situation, ecology, and nutritional status, and other countless variables are under suspicion.

Finding patterns in 58 counties of California can be helpful in further research of the question.

# Data Set

Preterm and Very Preterm Births by County, 1062 entries.

Table

Description automatically generated

## Data Transformation

What was done:

1. Names were lowered.
2. year, county, birth.type, total.births, and events were selected.
3. “None” values changed for 0.
4. Total results for California as a whole were deleted
5. Format was switched to wide from long, variable “events” to “preterm.births” and “very.preterm.births”
6. Rates were calculated

Final data set “data\_preterm”

'data.frame': 522 obs. of 9 variables:

$ year : num 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...

$ county : chr "Alameda" "Alpine" "Amador" "Butte" ...

$ total.births : int 19280 4 272 2447 346 338 12341 370 1617 16231 ...

$ normal.births : num 17658 4 247 2264 317 ...

$ preterm.births : num 1622 0 25 183 29 ...

$ very.preterm.births: num 259 0 0 31 0 0 152 0 21 262 ...

$ rate\_preterm : num 0.0841 0 0.0919 0.0748 0.0838 ...

$ rate\_very\_preterm : num 0.0134 0 0 0.0127 0 ...

## Exploratory Data Analysis

A screenshot of a computer

Description automatically generated with medium confidence

Counties are strongly differentiated.

This smoothscatter demonstrates yearly Total.births for all counties.

The highest portion, an enormous share, is that of LA County. There is no other county with more than 50k births a year.

There are a lot of counties with very low birth rate – deep blue zone.

We can clearly see that every county has its own statistics, but they share patterns inside some groups. Mostly, counties with higher birth rates have stable levels, and small counties have blank or fluctuating preterm birth rates.

A picture containing timeline

Description automatically generated

For very premature babies, the more severe pregnancy outcomes, patterns differ significantly more in different counties. Los Angeles has a consistent level, which, together with the highest total number of births, can be caused by better preparedness of the specialists and better ventilator equipment because extremely premature babies usually demand to individualize respiratory therapy.

Timeline

Description automatically generated

## A picture containing table Description automatically generatedTable Description automatically generated

These two plots demonstrate individual patterns (light green line) comparing to total state results (grey figure).

Here are plots using Smoothed conditional means and showing confidence interval around it for preterm\_rate (~7.5% of all births)

Chart

Description automatically generated

and very\_preterm\_rate (~0,75% of all births).

Chart, line chart

Description automatically generated

## Machine Learning

We will use K-Means cluster analysis to group similar observations.

To find the optimal number of clusters, we will investigate three methods: Gap statistics, Elbow, and Silhouette.

After all considerations, we take three centroids.

The size of clusters we get is 436, 77, 9.

## Chart, line chart Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

## Conclusion

This small data set approves the idea to find further regularities in county-level birth statistics.

We suppose there is a significant difference in preterm\_birth\_rate between the counties inside the clusters. In that case, the next step is to merge the socio-economic data or other datasets such as infections, race/ethnicity reports from governmental resources in order to find the correlation between unknown variables and outcomes of the pregnancies. Moreover, geographical factor was not investigated enough because of the problems with downloading maps without keys (chloropleth-family packages with maps)

## References

1. CALIFORNIA HEALTH AND HUMAN SERVICES OPEN DATA PORTAL

https://data.chhs.ca.gov/dataset/preterm-and-very-preterm-live-births

1. Machine learning and data science.
2. R for Spatial Statistics (Humboldt State University)
3. https://github.com/AMULETAnalytics/UCLAIntroDataScience