'US Birth Rate decline' – An attempt to find a cause

Project Report

By

Sai Yallapragada

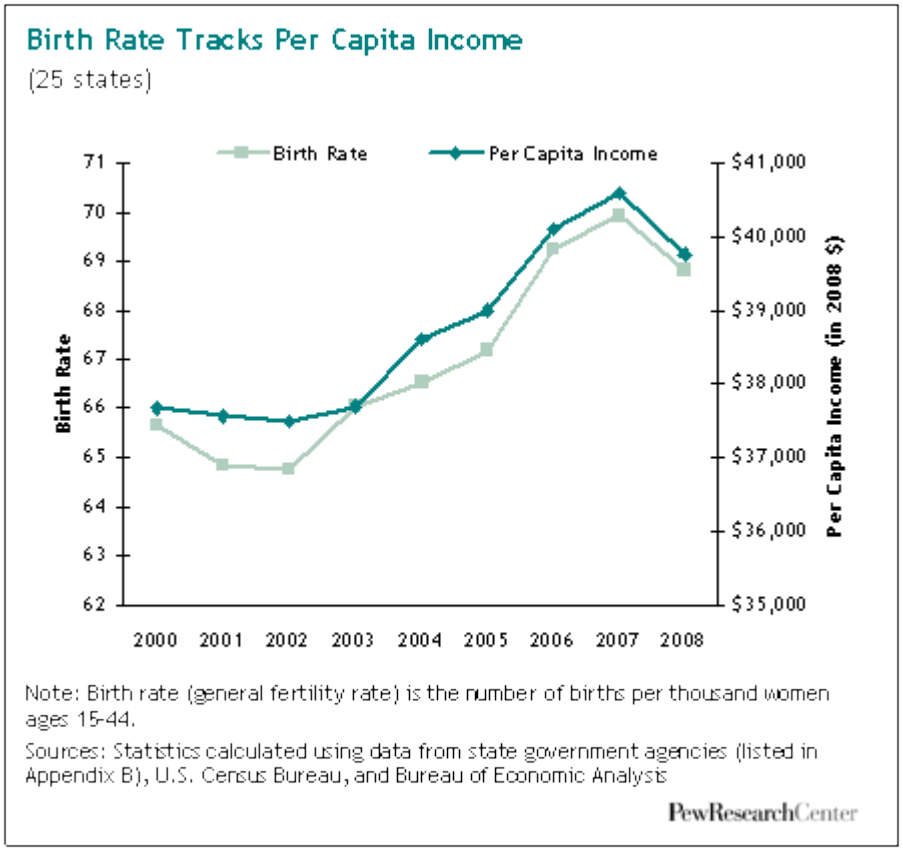
Executive Summary

This report was created as part of a final project for the Data Essentials course at University of Illinois, Springfield. The report summarizes a data model that was built in an attempt to see if there is any co-relation between the decline in Birth Rates in the US in last 82 years and common economic factors. The model considers economic parameters such as Stock market movements (Dow Jones industrial average), Inflation (Consumer Price Index), per capita Gross Income and Unemployment rate. Data was collected for all these variables and Random Forest was used to see if any of these variables have a correlation to the Gross Fertility Rate (Birth Rates) being lesser are greater than median Gross Fertility Rate in last 82 years. The result from the analysis shows that there is a strong link between Consumer Price Index and Gross Fertility rate movements.

Introduction

If you have followed the wide media coverage of the great economic recession of 2008-2009 you would have noticed that there are many faces of an economic recession. This recession has hit every part of our society like high unemployment, under employment, decline in home prices, home foreclosures, inflation, crumbling infrastructure, individual and corporate bankruptcies etc. But there is another face of the economic recession that has been not talked about a lot. It is the decline of birth rates in the U.S over the years, more so from 2008. Below is an article from Pew Research article that shows a decline in Birth rate from 2007 and 2008 and the decline is a sharp decline. The article finds a link between economic per capita income and birth rate.

http://www.pewsocialtrends.org/2010/04/06/us-birth-rate-decline-linked-to-recession/

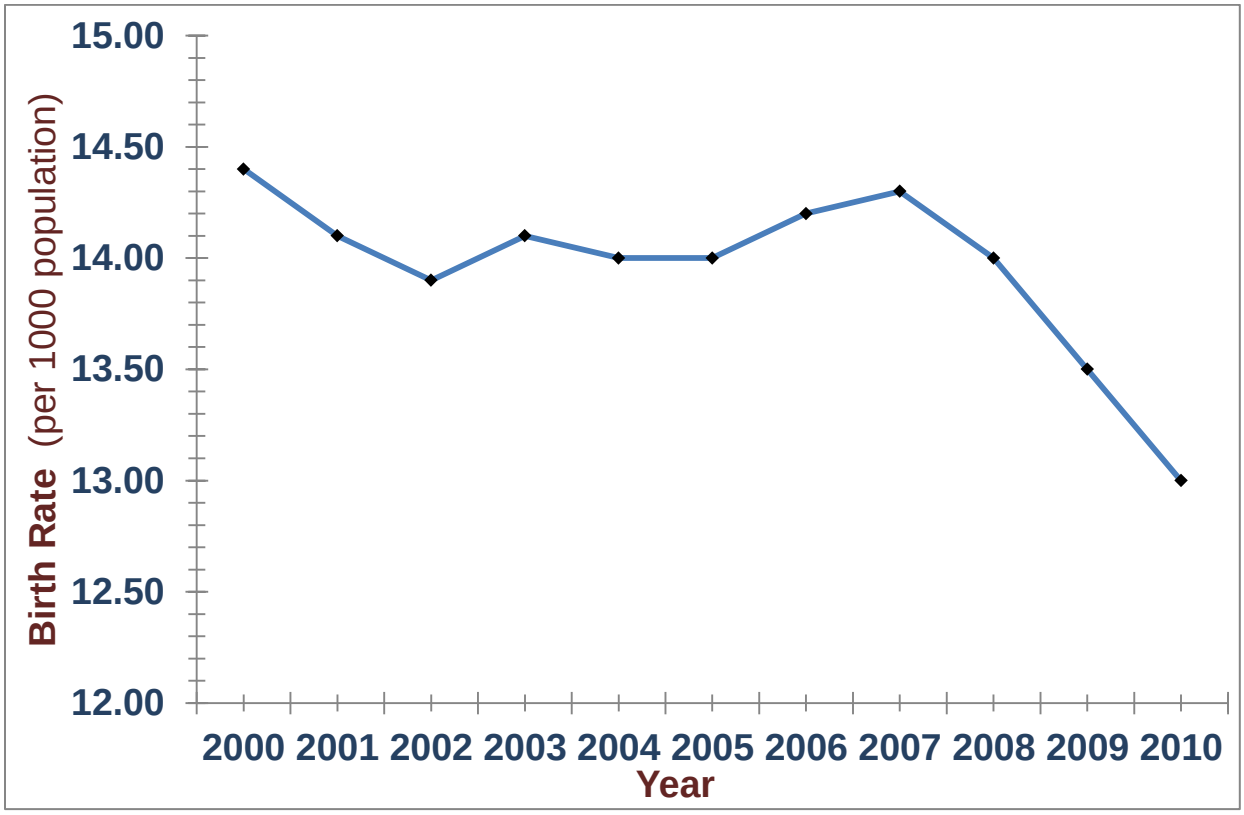


Birth rate is an important component of a modern world as current birth rates determine adult and working population numbers of the future which can translate into future tax payers that will fund the whole economy.

Population numbers are extremely important as they drive a country's competency. Population growth has both negatives and positives. Over population can put pressure on

existing infrastructure and it can lead to competition for resources such as water and food thereby driving the costs up. On the positives side, there will be tremendous human capital and high productivity. Declining population will lead to a smaller working force and an aging population. The consequences of this scenario are enormous such as lower number of entrepreneurs, lower tax revenues and unsustainable pressure on already over-burdened social programs. Overall, this will lead to lower competency on the world stage. Some economists have argued that the great recession of 2008-09 has not impacted countries like China and India because of their billion plus populations which exponentially increases consumer buying capacity thereby absorbing the impacts of an economic recession by consumer spending and overall human resources. Whether this argument is correct or not is a topic for a different day, but most people agree that population growth is an important fact in today's world.

In past few years there is a clear decline in the US birth rate. The birth rates from 2003-to 2007 have risen consistently and the birth rate started declining from 2008. Whether this decline is related to the economic recession is the subject of our investigation. To find a correlation between the birth rate and economic recession we have used historic US birth rates and historic US economic recessions. Please note that economic recession is defined as two consecutive quarters of decline in Gross Domestic Product **(GDP).** The source of historic birth rates in US census bureau and data source of the historic economic recessions is XXX. Here is a graphs that shows the sharp decline in the birthrate



Data Report

What may have happened in 2008 that caused the sharp decline in birth rates? There are many reasons for decline in Birth rates but this report considers various common economic indicators in a recession.

Hypothesis

- Our goal: to understand potential statistical relationship between Gross Fertility Rate(GFR) and various economic factors in the US.
- Hypothesis: Economic hardships, as measured by median per capita personal income, the unemployment rate, per capita GDP, DJIA value, and Consumer Price Index are associated with US Gross Fertility Rate (GFR) changes.

Independent variables (Predictors)

- Independent variables (Economic indicators)
- Per Capita Gross Income
- Unemployment Rate
- Per Capita GDP
- Dow Jones Industrial Average, annual % change
- Consumer Price Index (CPI)

Dependent Variable

- **Gross Fertility Rate (GFR)**

Data Collection

Data was collected for some of the common economic indexes during recession periods. These include – Recession periods, Dow Jones stock market movements over the years, Birth rates at national level, and Birth rates at state levels, US population, Income, Home foreclosure rates, per capita income, per capita GDP and Fertility rates. This data was cleaned up and scaled to appropriate levels.

Files

- CrossSectionData.xlsx
- Fredgraph.xlsx
- SampleSets.xlsx
- GFR_Analysis.xlsx

Data Analysis Process

After thorough analysis of the data, the unnecessary variables were removed. The remaining variables are –

- GFR
- US unemployment
- Per Capita Chained Income (at 2005 levels)
- Consumer Price Index
- Dow Jones movements as percentage loss/gain
- Median per capita Gross income (2007 dollars)

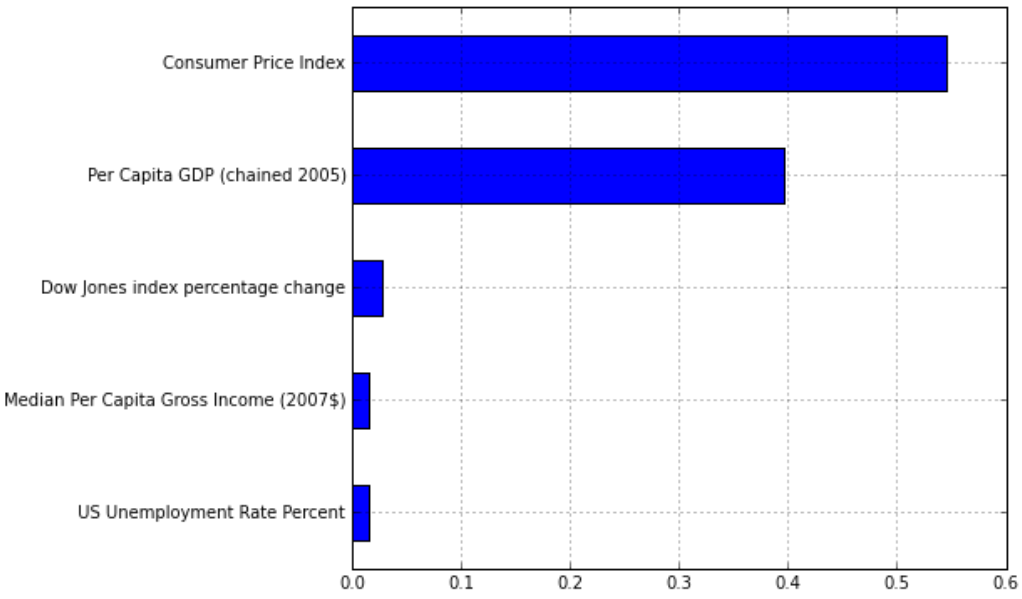
GFR_CleanedUp.csv has all the final variables with cleaned up data. The data was fed into the Pandas.

EDA

Exploratory Data analysis (EDA) was run on the provided data to see if it needs any clean up. The data that was brought into the model was already cleaned up to make sure there are no null values and data was appropriately used in Excel itself. After performing initial check on data, the dependent variable Gross Fertility Rate (GFR) was used as a dependent variable. And the "Year" variable was taken out as it does not serve purpose to finding a link between GFR and the economic indicators. Also, scaling the data to appropriate levels was performed. Now, we have a good data set.

Data Model

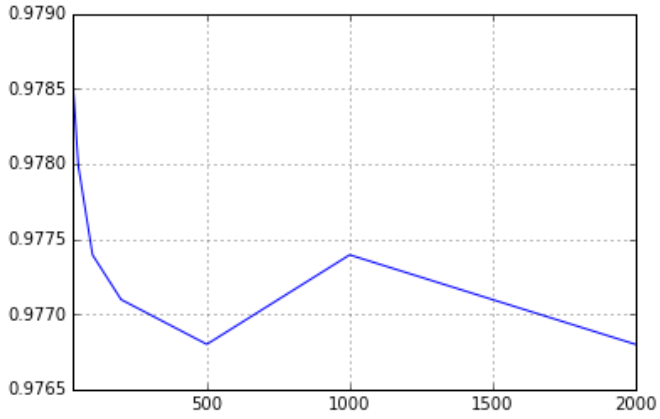
First, the data was split for a train and test situations and Linear Regression was applied to the data resulting in a R^2 value of 0.68, which is decent but far from good. Random Forest Regression was applied to the data. And then, Decision Tree regressor was applied which resulted in a R^2 of 0.98. Then Random Forest was applied with default values which gave us a R^2 of 1.0. At this stage, the dependent variable which is a linear value to a binary value because the goal of this project is to see if there is a link between the economic recession and Birth rate changes. The technique that was used for this is to calculate the mean GFR for last 82 years and anything above that was considered as a "good" GFR and anything that is below that number was considered "bad". The mean GFR for last 82 years is 76.7. The GFR column in the Data Frame was converted to 0 or 1 based on thus number. The Random Forester was applied again which resulted in an oob_score of 0.85 with a c-stat of 0.97. At this stage it is found that the most fitting economic indicator was the Consumer Price Index.



Data Model Tuning

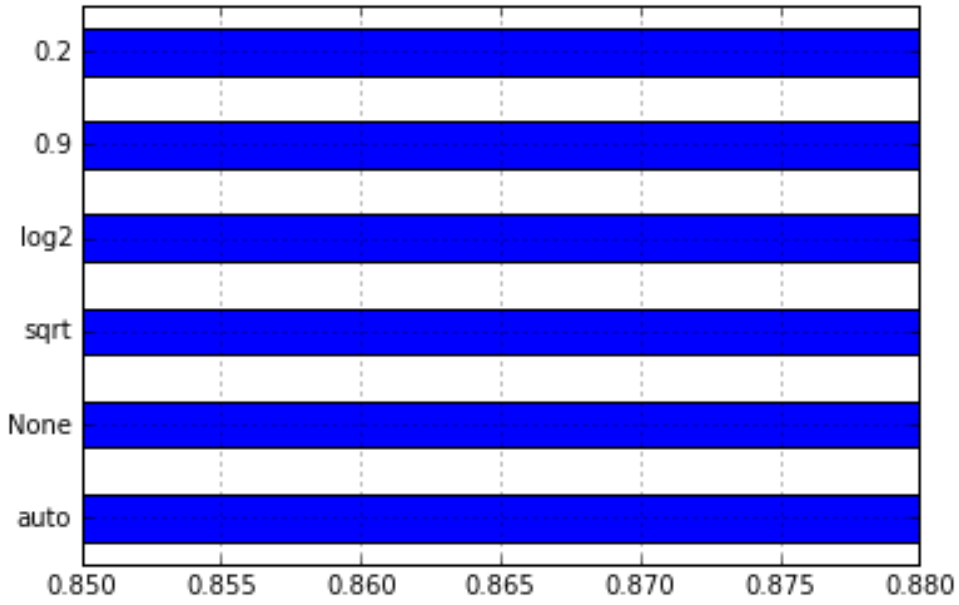
At this stage, the model was tried with various fits by looking for best fit for n-estimators, n-jobs, max_features to use min_sample leafs. And the results are as follows-

```
30 trees
C-stat:  0.978584176086

50 trees
C-stat:  0.977989292088

100 trees
C-stat:  0.97739440809

200 trees
C-stat:  0.977096966092

500 trees
C-stat:  0.976799524093

1000 trees
C-stat:  0.97739440809

2000 trees
C-stat:  0.976799524093
```



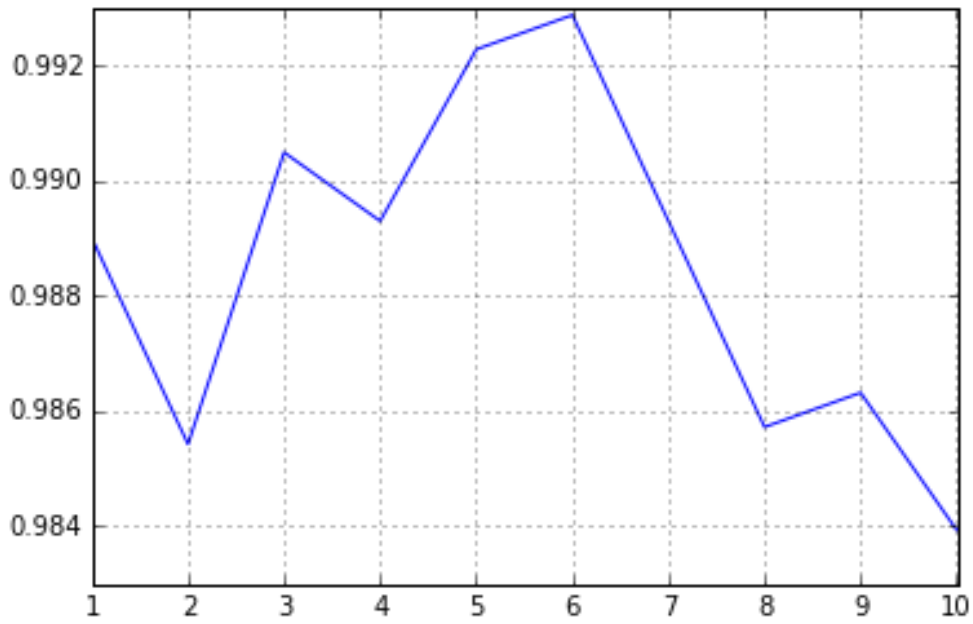

```
auto option
C-stat:  0.978584176086

None option
C-stat:  0.978584176086

sqrt option
```

```
C-stat:  0.979773944081

log2 option
C-stat:  0.979773944081

0.9 option
C-stat:  0.982153480071

0.2 option
C-stat:  0.988994646044
```




```
1 min samples
C-stat:  0.988994646044

2 min samples
C-stat:  0.985425342058

3 min samples
C-stat:  0.990481856038

4 min samples
C-stat:  0.989292088043

5 min samples
C-stat:  0.992266508031

6 min samples
C-stat:  0.992861392029
```

```
7 min samples
C-stat:  0.989292088043

8 min samples
C-stat:  0.985722784057

9 min samples
C-stat:  0.986317668055

10 min samples
C-stat:  0.983938132064
```



Final Model

Considering all the above results, the final model consisted of

- N_estimators – 30
- N_jobs - -1
- Max_features – 0.2
- Min_samples_leaf = 6

This resulted in a c-stat of 0.99.

Advanced Validation

Advanced Validation techniques were used on this final model to make sure that the model doesn't have over fitting problems. The accuracy score from the best random forest classifier was 0.94, which seems pretty good. And the classification report is as follows

	precision	recall	f1-score	support
0.0	0.91	1.00	0.95	10
1.0	1.00	0.86	0.92	7
avg / total	0.95	0.94	0.94	17

K-Force Cross Validation

K-Force Cross Validation was used on the analysis so far with the goal of validation the prediction about GFR using the economic parameters. The result is -

```
Score is 0.932500 +/- 0.097309
95 percent probability that if this experiment were repeated over and over
The average score would be between 0.835191 and 1.029809
```

The model doesn't seem ideal and it seems to have many problems including the predicted average score itself. But there could be many reasons for this including the data itself, data cleaning and unconsidered variables that might impact the GFR.

Conclusion

We can infer from the data analysis used here that there is some correlation between economic factors and the General Fertility Rate in the United States. **Most importantly we found that the Inflation rate (Consumer Price Index) has significant impact on the movement of the Gross Fertility Rate.** But it should be noted that this issue is complex and multiple dimensions of this issue need to be evaluated for better regression / prediction using better methods, better data, better data clean up and possibly more predictors needs to be used for better results. Examination of additional economic and social factors would be helpful in better understanding the relationship between birth rates and economic conditions. Some of these variables can be

- Ethnicity / age demographic
- Child Rearing costs
- Education levels
- Total Fertility Rate (TFR)
- Pure inflation
- Foreclosure data
- Social considerations
- State-level data
- Culture/Religion

By better understanding this phenomenon, a more complete picture can be painted of the long-term impacts of economic conditions historically, at present, and in the future.

Bibliography

- Pew Research – Social Trends
- http://stattrek.com/regression/linear-regression.aspx
- http://stattrek.com/regression/linear-regression.aspx
- http://en.wikipedia.org/wiki/Random_forest
- http://www.saedsayad.com/decision_tree_reg.htm
- http://en.wikipedia.org/wiki/Decision_tree_learning

Data Sources

- Center for Disease Control – National Center For Health Statistics www.cdc.gov/nchs
- Bureau of Labor Statistics www.bls.gov
- National Bureau of Economic Research www.nber.org
- US Census Bureau www.census.gov
- Bureau of Economic Analysis www.bea.gov
- Federal Reserve Economic Data http://research.stlouisfed.org/fred2/