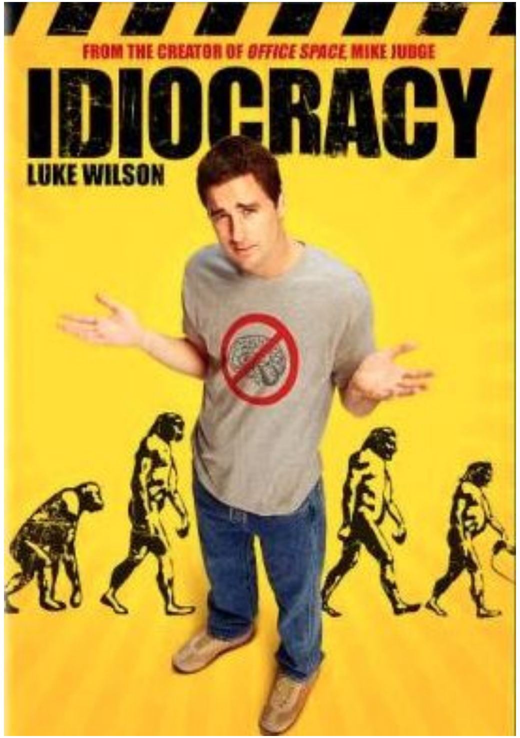
Understanding Factors That Affect Global Birth Rates

A Data Journey by Casey Hickcox, Jon Yarber, Luke Morris, and Blake Goehman

Outline

Background

The Original Idea



A Fruitful Dead End

Plenty of evidence to suggest the opposite:

- Intelligence is not hereditary!
- Humans are getting smarter
  - Flynn Effect

However:

- Yielded a wealth of population data
- Explains later findings

Tackling Global Population

- “There are too many people in the world” -Jon
- It took less than 50 years to double the world’s population
  - Almost 8 billion today
  - 4 billion in 1974
- Are there socioeconomic factors that affect this?
- If so, can they be manipulated to slow population growth?

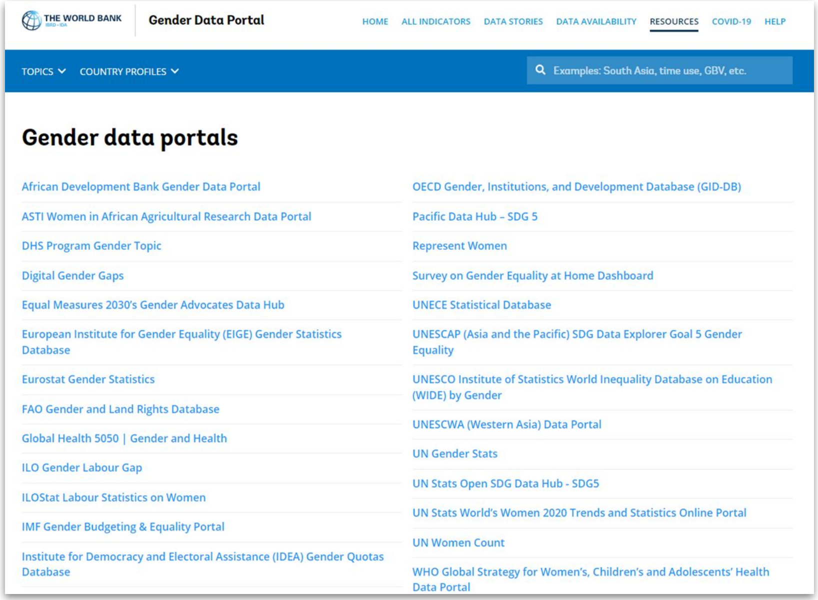
Initial Assumptions

1. There are specific economic factors which can be identified as mechanisms for controlling birth rates and therefore world population growth
    - Example: education
2. Some countries are experiencing an increase in birth rates while some are in decline
3. Birth rates could be modeled on a global scale: those factors that contribute to birth rates in Country A should also affect Country B

The Data

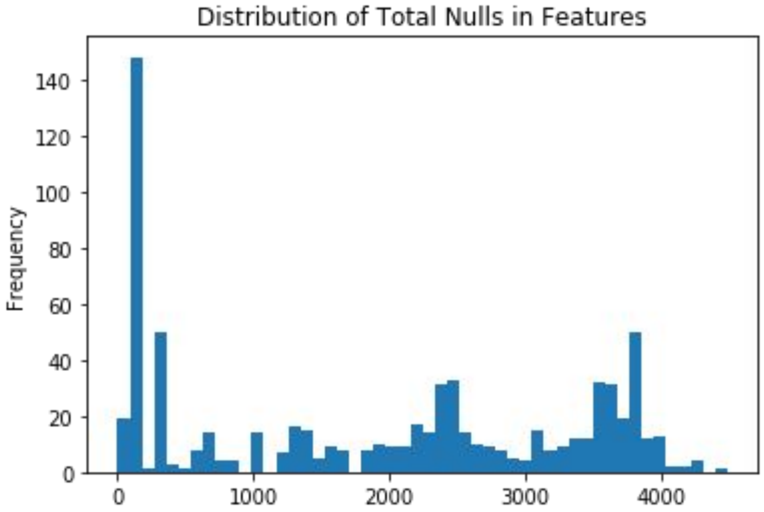
Creating the Data Frame

- World Bank
- 4 data sets:
  - Population Data
  - Gender Data
  - Health, Nutrition, and Population
  - World Development Indicators
- Combining the world bank data sets yielded:
  - 265 countries/territories
  - 1960-2020
  - 2,143 variables



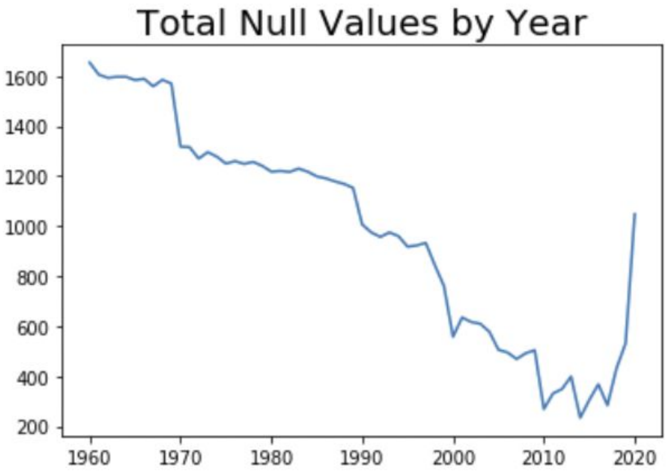
Issues With the Data Frame

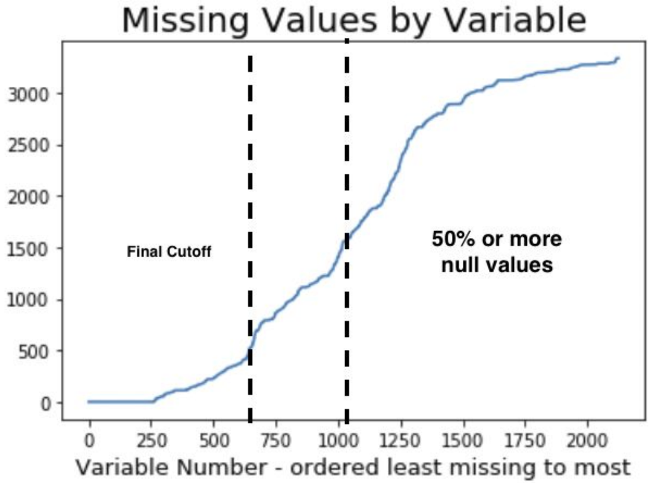
- 265 ‘countries’ in data frame but only 195 countries in the world
  - Economic groupings and territories
    - 'Fragile and conflict affected situations'
    - 'Central Europe and the Baltics'
- Missing data
  - 69.15% missing



Clean Up Process & Algorithm

- Joining on ISO 3166 alpha 3 codes eliminated demographic groupings and regions
  - 212 remained
  - Some territories are assigned alpha ISO 3166 3 codes - not really countries
    - Example: Cayman Islands - CYM
- Several iterations to find best process for reducing missing data. Final algorithm:
  - Reduce years
    - 1990-2018
      - 69% missing -> 58% missing
  - Reduce countries
    - Cut 81 countries, retained 95% of world population
      - 58% missing -> 51% missing
  - Reduce variables
    - Visual - knee
    - Reduced to 582 independent variables
      - 51% missing -> 4% missing
  - Impute remaining
- Stingy at first, but the algorithm created a process in which we could expand the data frame to include something that didn't make the cut if needed
  - Utilized often





	CountryName	NAs	TotalCells	%Missing
202	Liechtenstein	58555	63750	91.850980
203	Faroe Islands	58649	63750	91.998431
204	Monaco	58990	63750	92.533333
205	Turks and Caicos Islands	59384	63750	93.151373
206	American Samoa	59742	63750	93.712941
207	British Virgin Islands	59767	63750	93.752157
208	Gibraltar	60271	63750	94.542745
209	Northern Mariana Islands	60798	63750	95.369412
210	Isle of Man	61369	63750	96.265098
211	St. Martin (French part)	62720	63750	98.384314

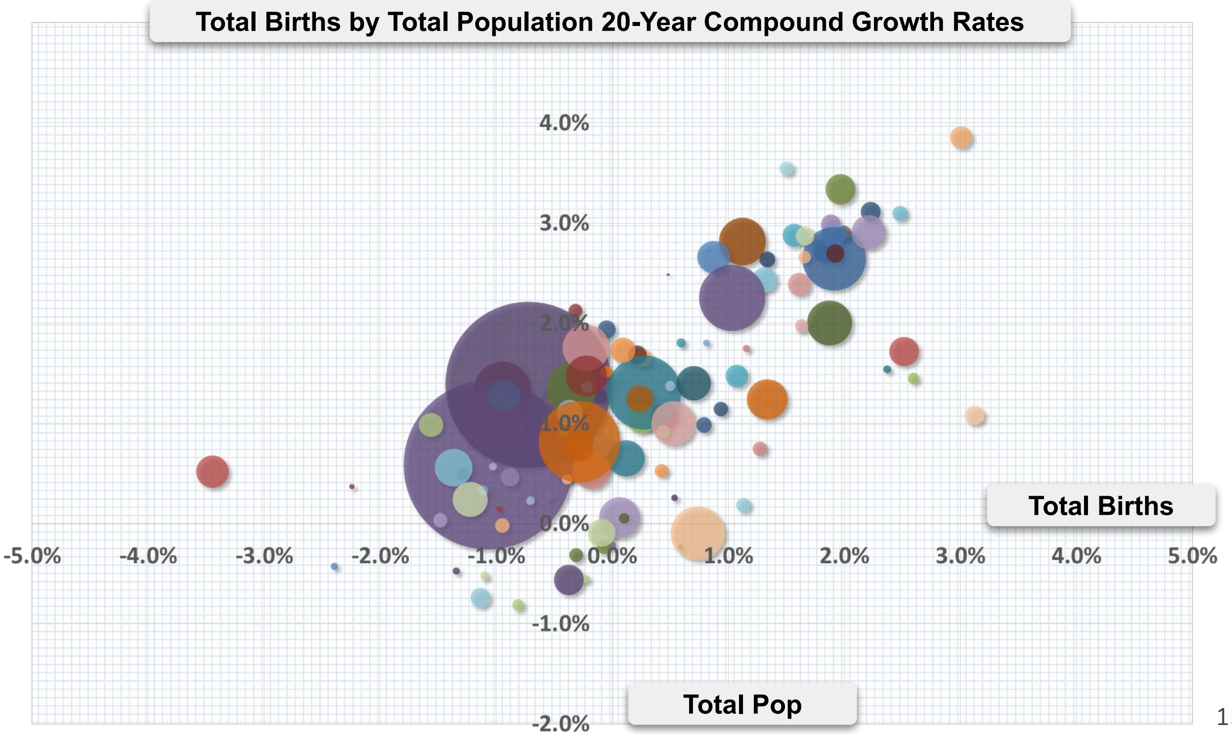
Groundwork & Initial Observations

Initial Questions While Entering Data Exploration

- What’s actually in the data frame? What are these variables?
- Are there any standout trends in population growth?
- How do demographics affect these trends?
- What’s our target variable?
- How do we deal with the potential (obvious) issue of collinearity?

Mid-Sized Countries Growing as Larger Countries Struggle to Replenish Population

- Mid-Sized countries have seen strong population growth trends over previous 20 years
- Larger countries have seen a decline in total births as population growth slows

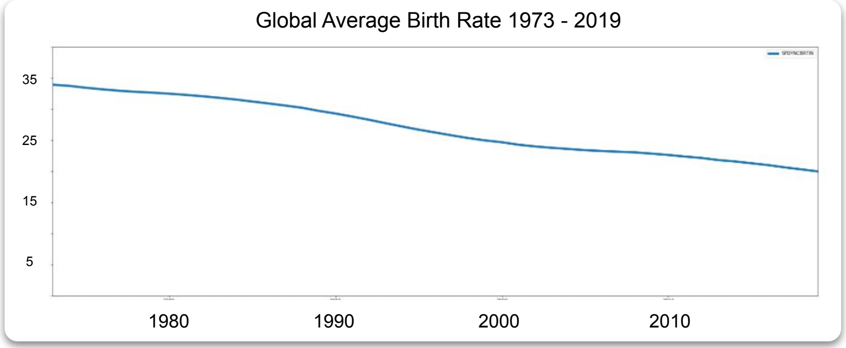


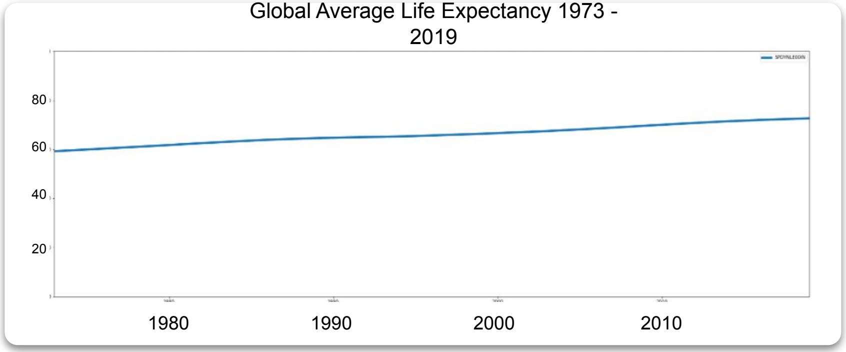
Standout Trends in Global Population Changes

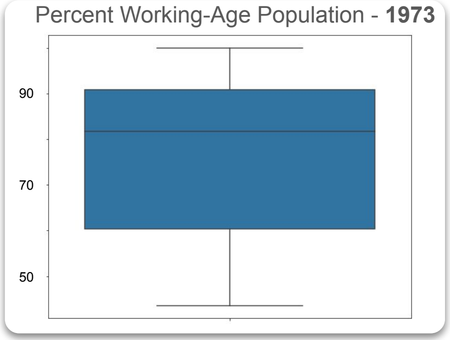
- Despite the rise in global population, birth rates and fertility rates have steadily decreased since 1960
- Fertility rate - number of children per woman - dropped from almost 6 in 1960 to a little less than 3 in 2020
- Rise in population more so a result of increase in life expectancy
  - Almost 70 in 1960 to almost 80 in 2020
- Leads to the issue of support for the elderly
- Future ‘healthy population’ controls may be necessary if only for this reason

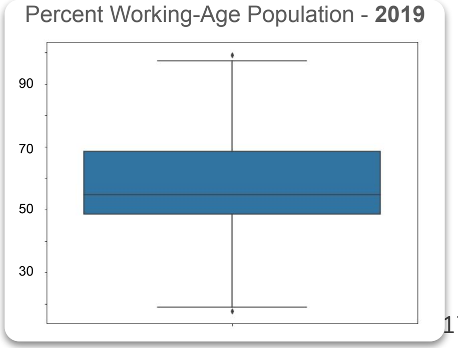
Working Age Population Pressured by Extended Life Expectancy

- As life expectancy rises and birth rates drop, working age populations shrink causing pressure on economies and social insurance programs.







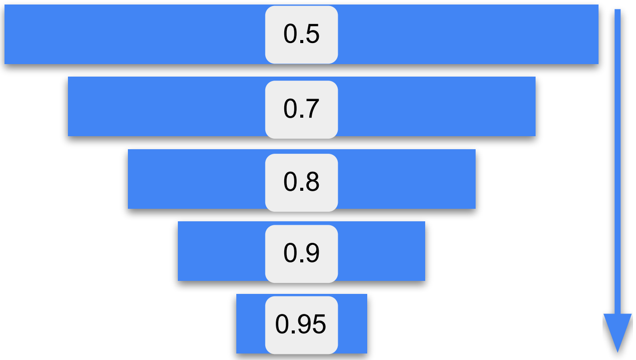


Evolution of the Target Variable

- ‘SPPOPGROW’: Total population growth
    - Problem: Immigration included
- ‘YOYORGGROW’: Manual calculation
    - Births - Deaths
    - Problem: Have no control over death rates (not ethically, anyway)
- The two combined above led to a manually calculated variable we added to the data frame: Immigration
- ‘SPDYNCBRTIN’: Birth rate, per 1000
    - Theoretically, this can be controlled to an extent and absolutely affects global population

Variable Relationships and Correlations

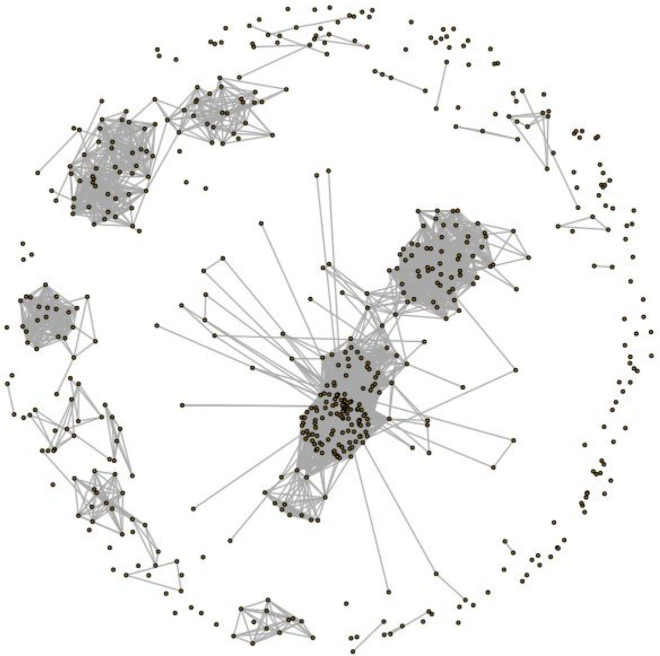
- Created a correlation data frame which contains the correlations between every variable in the data frame
  - Over 160,000 combinations
  - Derived from 580+ numeric variables
- Led to creation of custom correlation crawler
  - Allowed for understanding of ‘clusters’ of variables
  - Several were different measures for the same thing
    - Ex. 'GDP (current LCU)' & 'GDP: linked series (current LCU)'




```
def rabbit_hole(corr_df, variable, var2, var3, var4, var5, depth = 2, thresh = 0.5, pass_around = 0):

    x=-1 #Setting a counter - Could be set to zero if the x came after the changes, no difference
    if depth == 0: #Setting termination criteria
        #print("end of corr")
        return
```

Multicollinearity Created Complexity During Feature Selection

Features with Correlations Greater Than 0.9



- For the entire final data frame, over 5,500 correlations between variables were greater than 0.90
- Clustering of variables was attempted, but came back muddy.
- These clusters of highly-correlated variables caused complexity when entering the feature selection process as large swaths of valuable variables were very messy.

Feature Selection

Model Research & Selection

- Time series
- A lot of time and research went into finding which model made the most sense for us
- Final model of choice - panel data model using fixed effects
    - plm() in R built on lm() model and therefore had many of the same features and attributes as lm()
        - Other options were lacking those attributes and features

Feature Selection

- End goal at the time: find specific economic variables that could be manipulated to affect birth rates
    - No decomposition methods - PCA, non-negative matrix function, factor analysis, etc
- Knowing the model of choice allowed for a step-wise solution
    - Recursive forward elimination method evaluating variables on their p-values

When we finally ran the algorithm, it looked like this:

EXCEPT....

The Issues

- Over 100 variables being chosen
- Super high R^2 values (overfitting)
- Remove one variable and several others become insignificant
- Remove a variable from the selection process and the variable lineup changed

The issue of collinearity was bigger than we had anticipated it would be

Putting it to the test:

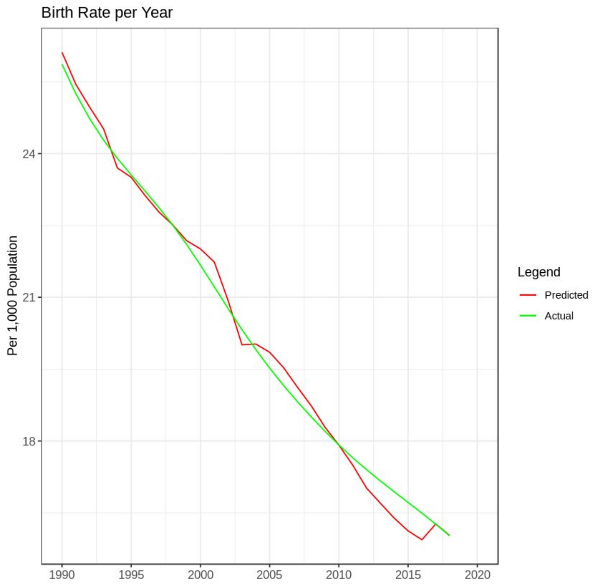
Modeling

Hypothesis Testing

- Stated:
  - Mechanisms exist with which interested parties could shape birthrates
- Assumed:
  - Birth rate is predictable with the data we have collected
  - Birth rate indicators are consistent enough that the train period is relevant to the test period
  - Birth rate trends exist on a large enough scale to be useful

Assumed Hypothesis 1: Data Modelability

- Models with high R^2
- Population cohort variable discussions
- Causation checked using the Granger Causality test usually showed two-way interactions

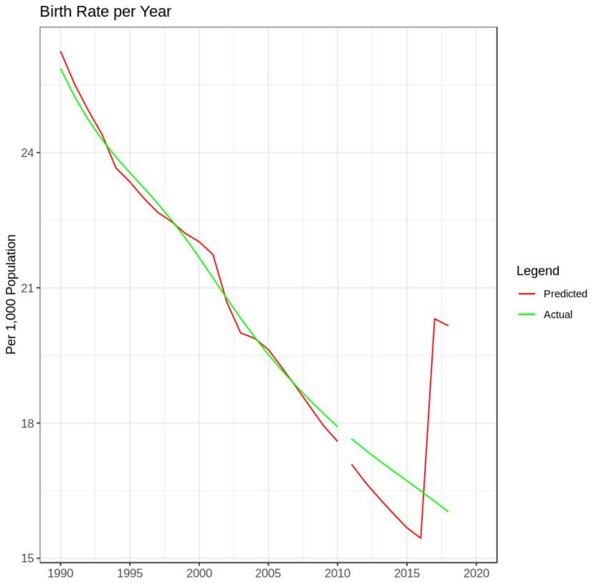


Assumed Hypothesis 2: Data Consistency

Intent: 1990-2010 Train / 2011-2018 Test

Problem: plm not generally used for predictive modeling

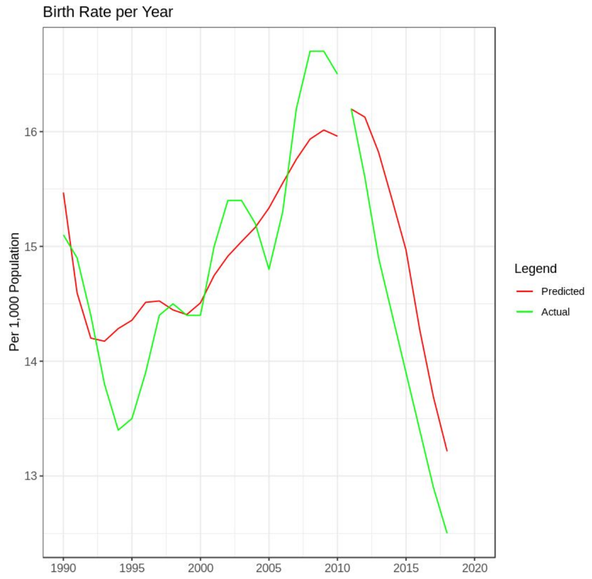
Solution: Use train period fixed effect values and coefficients to manually calculate the test period values



Assumed Hypothesis 2: Data Consistency

Realization: Population structure is actually an indispensable feature of our model

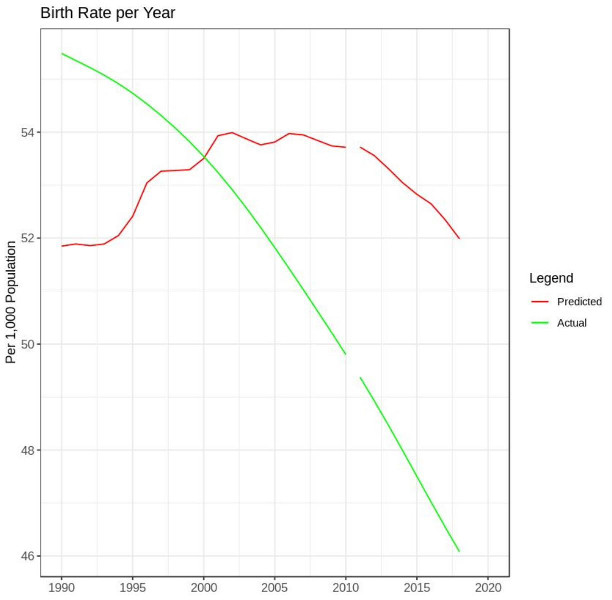
New approach: Test hypothesis with a pseudo-random-effects model built solely on population cohorts



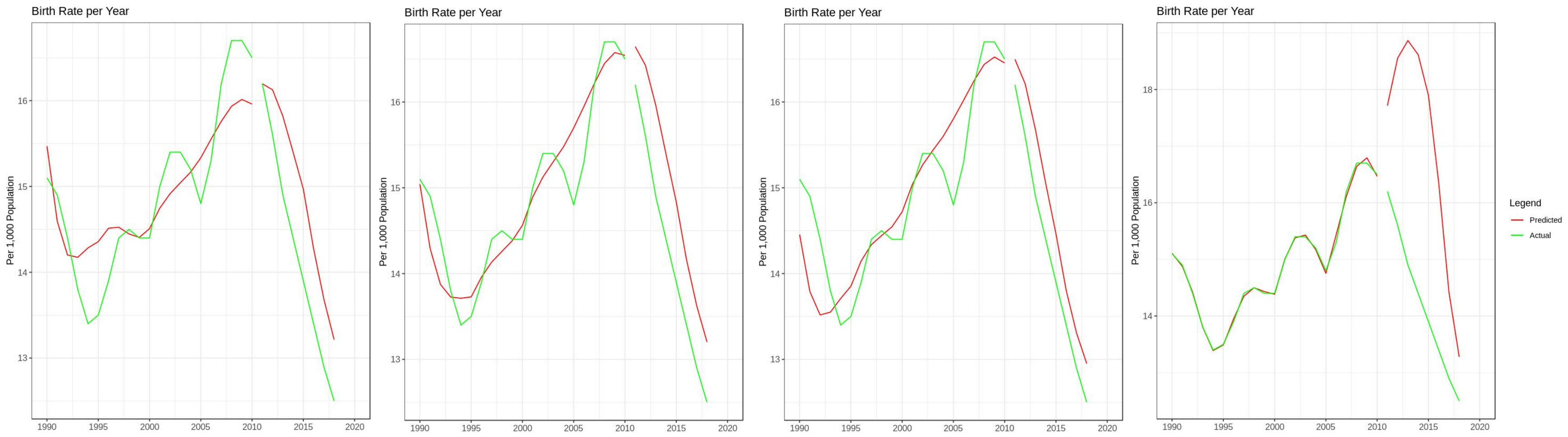
Assumed Hypothesis 3: Widespread trends

- Previous Indicators suggested sub-world modeling
    - Exploration stage showed differing trends
    - PCAs showed locale more important than year
    - Differences in modeling and correlations

- Tested using root mean square error to quantify performance



Resolving the Level of Focus: Ireland



World | Continent | Region | Country

Regionality Envisioned

Forward selection utilized again to produce:

- World model variables
  - Significance level .001 to reduce variable count
- Continental model variables
- Regional model results
  - Significantly fewer variables - as little as 5 in some cases
- Calculate RMSE for each country at each level and compare!

Findings

- World model produces worst results
  - Makes sense!
- Most countries perform better with the continental model
- Those countries that did not perform well on the continental model did better on the regional model and vice versa
- There are some countries that do not conform to any models

Let’s take a look!

Case Study - Africa

Africa

- Africa was used as an example of what can be done given time
- Meticulous selection of variables and model evaluation on a country level to find:
  - How certain countries affect the model
  - How certain variables affect the model
  - Which variables are important to this specific model
  - Which countries fit the model best, which don’t, and why

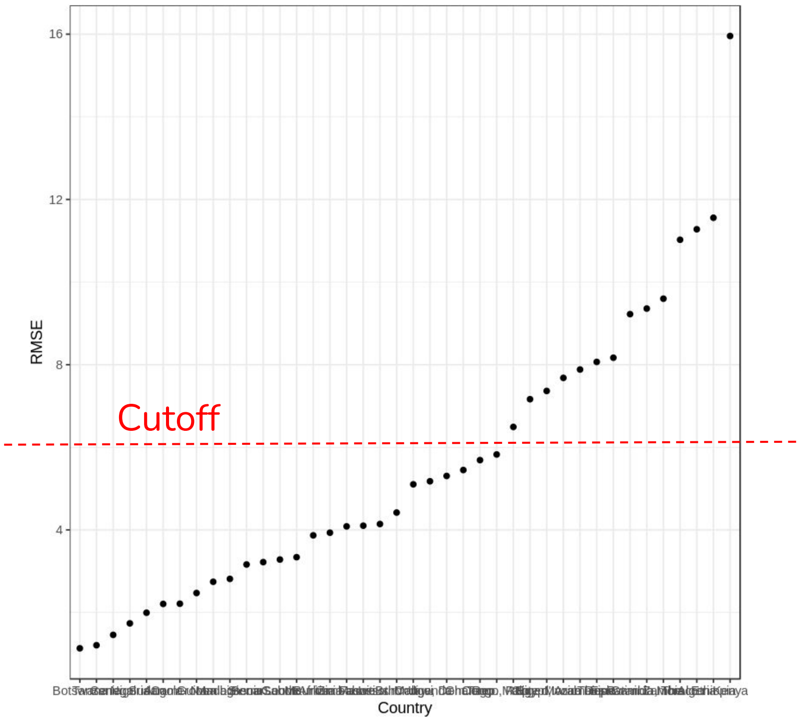
Feature Selection

- Variables sourced from feature selection methods previously mentioned
- Highest p value variable removed after each run until all features significant
- Mirroring variables removed
- Model rerun after every step

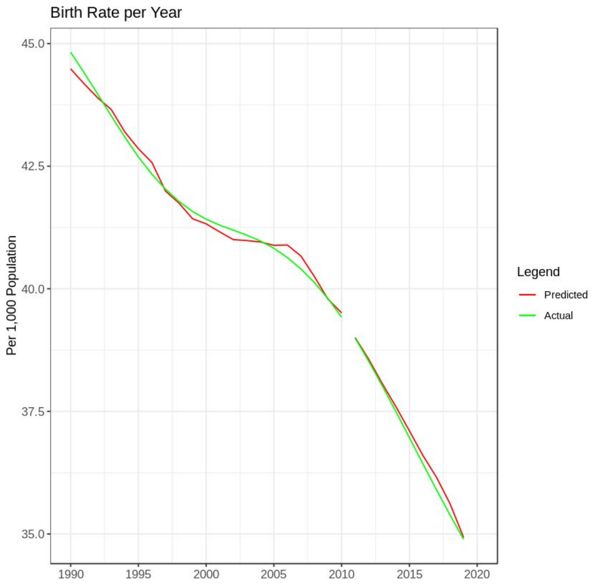
```
Oneway (individual) effect Within Model

Call:
plm(formula = eval(paste(dv, "~", ivstring)), data = train, effect = "individual",
    index = c("CountryName", "Year"), method = "fixed")

Balanced Panel: n = 26, T = 21, N = 546

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-0.087065 -0.021401 -0.002581  0.019766  0.087937

Coefficients:
               Estimate Std. Error  t-value  Pr(>|t|)
SESECENRR      0.0048575  0.0012833   3.7850 0.0001723 ***
SPPOP0509FE5Y -1.7013103  0.0737643 -23.0642 < 2.2e-16 ***
SPPOP65UPMAZS  0.9472651  0.1397193   6.7798 3.390e-11 ***
SPPOP65UPTOZS -2.2565416  0.1903003 -11.8578 < 2.2e-16 ***
SPPOP0509MA5Y  0.6184335  0.0527451  11.7249 < 2.2e-16 ***
SPDYNIMRTFEIN  0.0553603  0.0102737   5.3885 1.094e-07 ***
SPADOTFRT      0.1233721  0.0149910   8.2298 1.627e-15 ***
SPDYNTO65MAZS  0.2662688  0.0384246   6.9296 1.300e-11 ***
SPPOP1564FEZS -1.0806954  0.0420139 -25.7224 < 2.2e-16 ***
SPPOP1014MA5Y -0.4399312  0.0162645 -27.0485 < 2.2e-16 ***
SPDYNLE00MAIN -0.3687621  0.0443900  -8.3073 9.182e-16 ***
SPPOP4044MA5Y -0.0649313  0.0131218  -4.9484 1.023e-06 ***
SPURBGROW      0.0090094  0.0035764   2.5191 0.0120735 *
SESECDURS     -0.0101096  0.0035837  -2.8210 0.0049766 **
NECONGOVTZS   -0.0084894  0.0023658  -3.5884 0.0003653 ***
AGYLDCRELKG   -0.0318499  0.0124625  -2.5557 0.0108919 *
ENPOPDNST     -0.6095157  0.1394256  -4.3716 1.500e-05 ***
NVAGRTOTLCD    0.0538421  0.0158317   3.4009 0.0007252 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    10.88
Residual Sum of Squares: 0.46926
R-Squared:      0.95687
Adj. R-Squared: 0.95317
F-statistic: 618.704 on 18 and 502 DF, p-value: < 2.22e-16
```

Country Selection

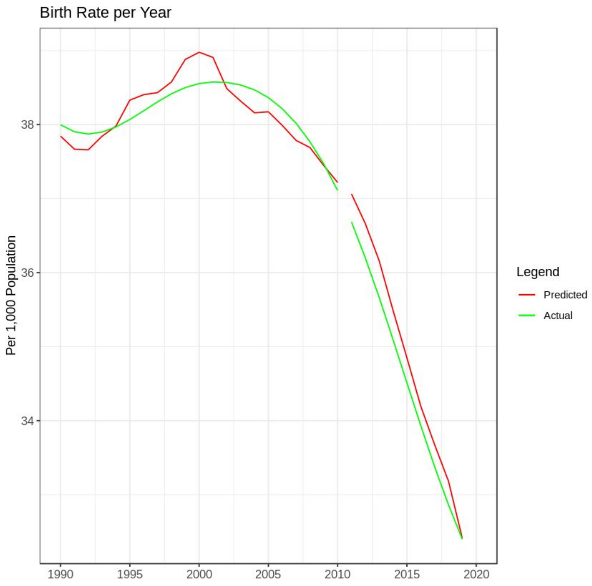
- 40 countries in Africa
- 26 made final model
- Countries removed primarily island nations, those with periods of conflict, and those with lasting foreign influence



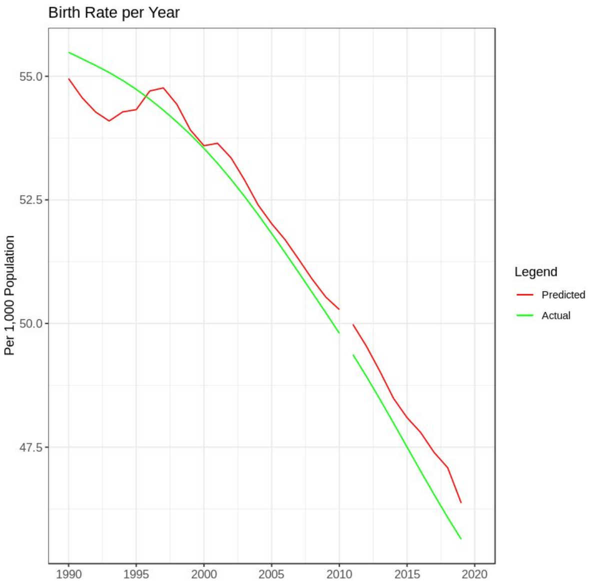
Final Model



Best



Middle



Worst

Final Discussion - Further Research

Potential Further Research, Part 1: What We Would Do Given a Few More Weeks

- Use clustering to group countries to model together
  - Regions we used are shown here
  - Using a clustering algorithm may group countries more effectively than geography alone
- Feature selection using correlation network
- Do what we did for Africa, for the whole world

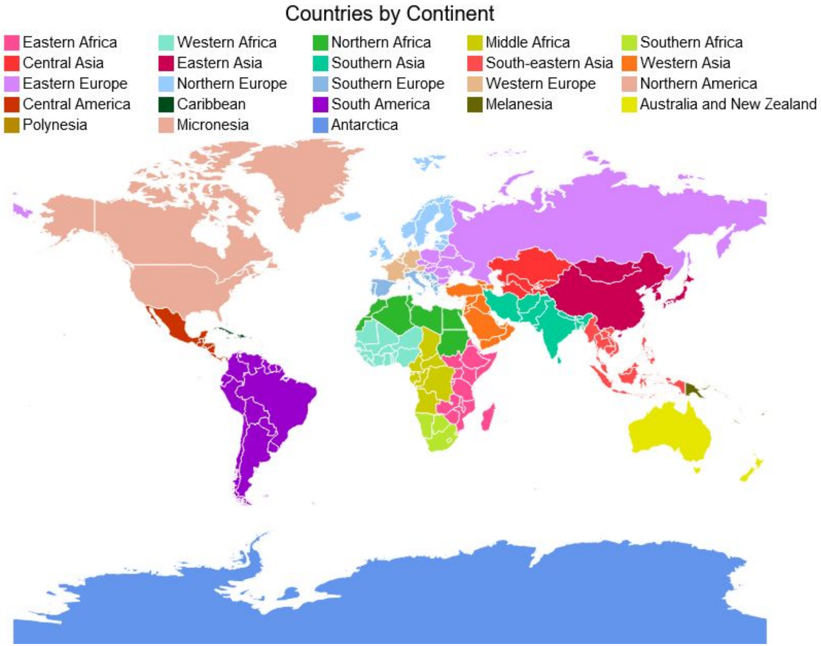


Image from StatisticsTimes.com

Potential Further Research, Part 2: Migration

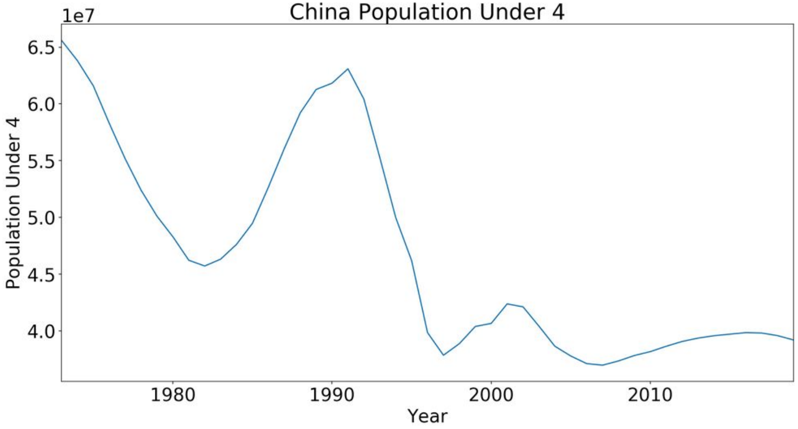
- Immigration is the only realistic alternative to increasing birth rate when it comes to fixing the demographic transition problem
    - Countries with low birth rates could take more immigrants
    - Migration allows countries to quickly ‘reset’ their demographic breakdown
- Migration is hard to keep track of
    - Data is not comprehensive
- Unintended consequences
    - Mistreatment - migrant slaves in United Arab Emirates
    - May push the problem to poorer countries

Potential Further Research, Part 3: Other Methods and Models

- Neural network on panel data
- Cascaded/multi-stage regression (Outputs of country model are inputs of region/continent/world model)
  - The first stage is not limited to panel data models
- Stacking (training multiple models and making predictions by combining the model outputs)
- More sophisticated data carpentry
  - We had a lot of missing data and didn’t want to impute too much
  - We could use domain knowledge to pick specific variables that are more likely to impact birth rate, despite missing many values

Potential Further Research, Part 4: Additional Paths of Investigation

- Add data from specific countries to be used on first stage of multi-stage regression
  - Annual statistics from government agencies
- In-depth analysis of China's one child policy
  - How effective was it?
  - What are the lasting effects?
- Public service announcements
  - Less oppressive than a legal requirement
  - Would they work?



Questions?

Appendix

Variable Definitions

Created a function that would take a variable or an array of variables and produce a corresponding array of their definitions.

```
def ind_def_lookup(indicatorarray):

    defs = []

    indicatordict = pd.read_csv('/dsa/home/jyymn/jupyter/sp22Capstone_01_Group03/GroupProducts/Milestone1/Indicator_Dic

    indicatordict.columns = ["IndicatorCode", "IndicatorName"]

    for ind in indicatorarray:
        if ind not in indicatordict["IndicatorCode"].unique():
            defs.append(ind)
        else:
            defs.append(indicatordict[indicatordict["IndicatorCode"] == ind]["IndicatorName"].values[0])

    return defs
```