Solutions for the declining birth rate in the ROK

Milestone One

Ross L. Kim (Schreck)

Bellevue University

DSC520 Statistics for Data Science

Professor Denton

2023.10.24

Introduction

For the term project, I hope to delve into some demographic issues of the Republic of Korea where I am currently living. South Korea is a flourishing democracy with the 13th largest economy in the world and 3rd largest in Asia. South Korea experienced a baby boom in the 1970s and 1980s which rendered the metropolitan areas overpopulated; South Korea's infrastructure was not developed enough at the time. After rebuilding from the aftermath of the Korean war, the South Korean government took an 'economy first' initiative. Despite being a democracy, conglomerates (Chaebol) were appointed specific industries in which they could maneuver, and in most ways, each conglomerate, to this day, is constructed and functions like a monarchy, where the CEO is king, and nepotism is rampant. The government played a major role in dissuading its population from having more than two children in the 80s. Since then, the birth rate has plummeted and is now the lowest in the world. This will affect society in all facets. The trend is very similar to that which Japan experienced in the late late 90s and two-thousands, and China is projected to follow Korea in years to come. This has serious implications on East Asia as a whole in the near future potentially directly affecting over a fifth of the world's population.

Research Questions

1. What problems could Korea encounter if this trend continues?
2. How much has the rising divorce rate been a major contributor? How can the population be dissuaded from having a relatively conservative attitude toward the practice?
3. How much have housing prices been a major contributor? How is it contributing and how can it be solved?
4. How has the cost of living been a major contributor? What economic implications or repercussions are there should regulations be implemented?
5. How has strict social adherence to the importance of bloodline contributed? Is unwillingness to domestically adopt still an issue today?
6. How has suicide contributed? The ROK has the highest suicide and lowest birth rates in the world. What correlations are there between the two?
7. How can the ROK's economic and demographic trends and progression over time compare to that of other nations (particularly Japan)?
8. How could future reunification have an impact?
9. How can this problem be addressed at an administrative level?

Approach

I plan to take into account as many variables contributing to this issue as possible. I plan to make a comprehensive comparison of societal conditions since the Korean war into modernity. This approach will address the main culprits, how they've evolved socially over the decades, and how they can either be reversed or what counter measures could be considered.

How my research addresses the problem

In my approach, I will deep dive into the variables that are considered contributing factors, beginning with population itself. I will compare the population trends of the ROK from pre modernization into modernity. With regard to population, the trends of the ROK will be assessed and compared to those of similar countries. Similarities and differences will be evaluated as a way to distinguish comparable and uniquely regional insights. This research will undoubtedly isolate the variable of interest to conduct further research. My research of the ROK's GDP, industrial growth, purchasing power, and income levels, I hope to shed light on how these economic variables correlate with one another, and if so, explore its comparison and correlation with the population trends. With this research, I hope to realize the economic implications and considerations. Other demographic considerations should be addressed such as population disbursement throughout the peninsula, industrial shifts, and other social tendencies among populations throughout the peninsula. Real estate prices undoubtedly play a large role; researching and understanding trends in price, types of dwellings, types of leases, etc. will shed much light on a likely key factor in the issue. Researching the suicide rate, which the ROK is ranked number one in highest rate in the world, will help to understand what facets in daily life are the most troublesome to the individual, and how much this affects the overall population.

Data

1. Global Population Trends - statistics for all nations in the following categories: total population, urban population, rural population, population density, life expectancy, birth rate, death rate, fertility rate, infant mortality rate, growth rate
2. ROK Income Levels - economic growth and sustainability of the ROK; income levels of the ROK past to present
3. ROK Demographics - statistics for the ROK in the following categories: birth rate, death rate, marriage rate, divorce rate
4. ROK Real Estate Prices - locations, households, building age, building grade, lowest price, highest prices, average prices
5. WHO Suicide Statistics - suicide statistics for every nation in the following categories: country, year, gender, age, number of suicides to overall population

Required Packages

1. ggplot2
2. readxl
3. dplyr
4. datasets
5. Purr
6. tidyverse

Plots and Tables

I plan to use many plots and histograms for this research project. Due to many data sets' chronological nature, I will especially need ggplot for the visualization of histograms and line graphs to show things like the many trends and rates that I will need to visualize. For example, population trends, birth rates, death rates, life expectancy, income levels, costs, etc. will all require histograms to show change over time. I will also need scatter and box plots when transforming data and making comparisons. For example, I will likely need to transform data that has missing variables and outliers; I will also need to compare many inter-national variables such as population densities, birth rates, etc.; I will also need to compare real estate statistics from region to region.

Questions for Future Steps

1. How can this research be used or built upon to forecast the population's current trajectory?
2. How can this data be used to persuade known unethical social norms (eg. bloodline, intolerance to divorce, etc.)?
3. How can this research be used or built upon to forecast housing price trends? Based on the data, what measures seem reasonable in order to combat this issue?
4. How can this research be used or built upon to forecast costs of living?
5. How can this research be used or built upon to forecast the suicide rate? Assuming many issues are realized through this research, what further investigation should be taken to tackle this extremely significant issue?
6. How did a comparison to other nations help in realizing what can be anticipated in the future? Based on these researched cases (eg. Japan, China), What would be reasonable to assume, and what relevant questions should be answered in follow up research?
7. How can this research give insight into current regulations and policies' influence in the present day circumstance? Based on all comparisons and forecasts, what additional regulations and policies should be considered at the administrative level?

Solutions for the declining birth rate in the ROK

Milestone Two

Ross L. Kim (Schreck)

Bellevue University

DSC520 Statistics for Data Science

Professor Denton

2023.11.06

How to import and clean my data

I imported the following datasets in csv format:

1. apt_seoul.csv
2. population(2016-2022).csv
3. real_estate_seoul.csv
4. ROK_demographics_2000-2022.csv
5. who_suicide_statistics.csv
6. ROK_income_welfare.csv

For dataset 01 (dt01), I executed the following steps for cleaning:

1. Removal of empty rows - 4,544 rows remaining
2. Removal of empty columns - 11 columns remaining
3. Replaced NA with median value - affecting rows 4, 5, 6, 7, 10
4. Omitted NA - 0 NA remaining
5. Renamed working variable to ‘dt01_tf’

For dataset 02 (dt02), I executed the following steps for cleaning:

1. Removal of empty rows - 1,073 rows remaining
2. Removal of empty columns - 12 columns remaining
3. Renamed working variable to ‘dt02_tf’

For dataset 03 (dt03), I executed the following steps for cleaning:

1. Change of column 7 name to ‘square meters’
2. Change of column 8 name to 평 (pyung)
3. Removal of empty rows - 4,021 rows remaining
4. Removal of empty columns - 11 columns remaining
5. Replaced NA with median value - affecting rows 8, 9, 10
6. Omitted NA - 0 NA remaining
7. Renamed working variable to ‘dt03_tf’

For dataset 04 (dt04), I executed the following steps for cleaning:

1. Removal of empty rows - 4,860 rows remaining
2. Removal of empty columns - 12 columns remaining
3. Replaced NA with median value - affecting rows 2, 3, 4, 5, 6, 7, 8, 9, 10
4. Omitted NA - 0 NA remaining
5. Renamed working variable to ‘dt04_tf’

For dataset 05 (dt05), I executed the following steps for cleaning:

1. Removal of empty rows - 43,776 rows remaining
2. Removal of empty columns - 6 columns remaining
3. Replaced NA with median value - affecting rows 4, 5
4. Omitted NA - 0 NA remaining
5. Renamed working variable to ‘dt05_tf’

For dataset 06 (dt06), I executed the following steps for cleaning:

1. Removal of empty rows - 92,857 rows remaining
2. Removal of empty columns - 14 columns remaining
3. Replaced NA with median value - affecting rows 11, 12, 13
4. Renamed working variable to ‘dt06_tf’

Final data sets

The main dataset that I will be using is ROK demographics (ROK demographics 2000-2022.csv). With 12 columns, it will likely be the most useful source for my research. Columns include the following: date, region, birth, birth rate, death, death rate, divorce, divorce rate, marriage, marriage rate, natural growth, and natural growth rate. I will need to confirm the units according to the categories and assign them in the next step. Likely, all variables in this dataset will be relevant.

The dataset of apartment costs in Seoul (apt_seoul.csv) will be a source to extract housing prices in the most densely populated metropolitan area. This set contains data of every major apartment complex in all districts and regions in the Seoul area. Likely at least two variables will be relevant: average sales and price per area.

The dataset containing world populations (population(2016-2022).csv) will be a source used to compare internationally. This dataset includes population trends, birth rates, death rates, life expectancy etc. for every nation. Likely several variables will be relevant: birth rate, growth rate, and death rate for several nations that are comparable to the ROK (Japan, UAE, China, etc.)

The dataset containing land values in Seoul (real_estate_seoul.csv) will be a source to compare property values with infrastructural development. This includes the coordinates, number

of households, land value, sizes, and sales prices over periods of time. Variables of interest will likely be sales prices over periods of time.

The dataset of suicide statistics (who_suicide_statistics.csv) will be a source used to compare rates internationally. Korea has the highest suicide rate whilst having the lowest birth rate, so variables from this dataset will likely be relevant. Relevant categories include country, rate, age, gender, and years.

The dataset of income (ROK_income_welfare.csv) will be a source used to compare individual incomes with cost of living particularly. Relevant categories include year, income, gender, education level, marital status, and occupation.

Future Steps

1. Need to confirm the units for all variables I plan to use.
2. With two similar datasets (apartments in Seoul and real estate in Seoul), how are they comparable, and which overlapping data, if any, should I omit?
3. With the majority of the ROK's population living in the Seoul area, is the use of primarily housing data from the Seoul metropolitan area necessary?
4. Is suicide data impactful enough on the overall population to include in this research? It is undoubtedly a major social issue, but is it relevant regarding population decline?
5. How can I take variables from the multiple datasets and compile them using a normalizing method? How will I need to make adjustments accordingly?
6. I may need to consider ways to romanize some data that is in Korean in case visualization for a foreign audience is needed.

Information that is not self evident

As mentioned previously, some units will need to be assigned to variables as they are ambiguous in their current forms. In the comparison of international statistics to the ROK, it is not entirely clear which nations would be of best comparison. I will need to conduct further research into what countries share similarities in demographics and trends in recent history to determine this. Some datasets, such as housing data, only go back as far as 2017 impeding the likelihood of making an accurate forecast of future housing costs, which was set in milestone one. The meaning of the growth rate variable is ambiguous and needs to be clarified.

Ways to look at the data

As mentioned previously, relevant variables will be extracted from some datasets and combined with the main dataset (ROK demographics). I will begin by looking at the past data to form a conclusion of how events lead up to the current situation. I will need to analyze the trajectories of several continuous variables in order to achieve this. For example, in what time periods the ROK encountered rapid population growth compared to time periods when the population declined. What other variables had statistical significance during these time periods?

I will then analyze the ‘how’ by identifying the patterns within the data. Finding the cause of these outcomes will help to develop a better understanding of countermeasures that could be suggested in the future. I hope to develop accurate predictive models for the relevant variables by taking into account every possible contributor while focusing on the main dependent variable, the declining birth rate.

Ways to slice and dice the data

There are many categories to draw from. I will extract three dimensions (average sales, area, and price per area) from the apartment dataset; seven dimensions (relevant country, year, total population, life expectancy, birth rate, death rate, and growth rate) from the international populations dataset; one dimension (average sales) from the housing dataset; all 12 dimensions (date, region, birth, birth rate, death, death rate, divorce, divorce rate, marriage, marriage rate, natural growth, and natural growth rate) from the ROK demographics dataset; four dimensions (relevant country, year, age, suicide rate) from the suicides dataset; and six dimensions (year, income, gender, education level, marital status, and occupation) from the income dataset. I hope to normalize a few key quantifiable measures (population and birth rate) as independent variables and adjust dependent variables for conformity and consistency while maintaining a minimal margin of error. I will use functions within the dplr package to slice and arrange data to accomplish this. When making predictive models based on the current data, I will use population and birth rate as dependent variables.

Summarizing data to answer key questions

The following are the questions I asked in milestone one:

1. How can this research be used or built upon to forecast the population's current trajectory?
2. How can this data be used to persuade known unethical social norms (eg. bloodline, intolerance to divorce, etc.)?
3. How can this research be used or built upon to forecast housing price trends? Based on the data, what measures seem reasonable in order to combat this issue?

4. How can this research be used or built upon to forecast costs of living?
5. How can this research be used or built upon to forecast the suicide rate? Assuming many issues are realized through this research, what further investigation should be taken to tackle this extremely significant issue?
6. How did a comparison to other nations help in realizing what can be anticipated in the future? Based on these researched cases (eg. Japan, China), What would be reasonable to assume, and what relevant questions should be answered in follow up research?
7. How can this research give insight into current regulations and policies' influence in the present day circumstance? Based on all comparisons and forecasts, what additional regulations and policies should be considered at the administrative level?

I will use measures of correlation to make forecasts for the above questions. Multivariate insights can be used to summarize the correlation among, for example, the population's future trajectory. An instance would be the measure and comparison of correlations among birth rate, marriage rate, and life expectancy. This would likely result in negative or positive correlations. For example, it could be assumed that as life expectancy increases, then the birth rate decreases. I will use correlations to investigate and correlation of housing price trends, mariage rate which, I'm hoping, will lend further insights into the birth rate. Correlation can also be applied to cost-of-living forecasts, perhaps by comparing marriage rate, birth rate, and housing prices. Perhaps it can also be applied to variables of the same dimensions of a similar nation, such as Japan. A country which has already undergone its social and economic transformations that the ROK is just beginning. Both nations also have similar cultural and linguistic traits, demographics, social constructs, values, and a somewhat shared history.

I will summarize the central tendencies of periods of time to make comparisons with economic developmental stages. For example, what was the mean of the annual birth rate from 1963-1979 during the Park Jeonghui administration, when the ROK experienced unprecedented economic growth? What was the mean for the same variable following the financial crisis of 2008? I will also summarize the variability of the data to answer questions. Variance could be relevant when considering gender and income of individuals. It should be assumed that while gender disparity is still an issue today, it would have been far more severe in the past.

Plots and Tables

As mentioned in milestone one, I plan to use many plots and histograms for this research project. Regarding packages, I will need ggplot2 and ggstatplot. Also as mentioned before, due to the chronological nature of my data, I will likely use histograms and line charts the most. They will be ideal in visualizing trends and patterns for comparisons.

Machine Learning

I will likely need to research and try to acquire the technical knowledge in the endeavor of applying machine learning to my research. It's quite obvious that creating an algorithm based on my current data could be a key tool in developing an accurate predictive model. I could use linear regression models to estimate upcoming trends. Some possibly relevant packages that I'm not already using to consider could be data.table and caret for regression training.

Filtering, reducing, and arranging data

Reduction of data will be essential as I am compiling variables from multiple datasets. As mentioned above, I will use only four to five total variables from the apartment and real estate datasets. Of the six datasets that I am using, there are 66 total variables; I will be using fewer than 30. After the necessary reduction and compilation of variables, I hope to have fewer than 15. Filtering certain rows will also be necessary. Such as years that are irrelevant.

Solutions for the declining birth rate in the ROK

Milestone Three

Ross L. Kim (Schreck)

Bellevue University

DSC520 Statistics for Data Science

Professor Denton

2023.11.18

Introduction

For the term project, I have used multiple variables from six different data sets to answer the questions previously addressed. A few of the questions were unanswerable as I did not have the sufficient data. The primary goal of conducting this research was to determine the causes of the declining birth rate of the Republic of Korea. A summary of the main variables used for extraction starts with birth rate. Divorce rate, marriage rate, housing prices, and income were also widely used. I have had to adjust some initial questions from previous milestones due to insufficient data. I have narrowed them down to eight problems that were fully or partially addressed as stated in the section below, 'Problems Addressed'.

Problems Addressed

1. Current and future trajectories of key variables such as birth rate and natural growth rate.
2. The impact divorce rate has on birth rate.
3. The impact of housing prices on the birth rate.
4. The cost of living and its impact on the birth rate.
5. The effect of social adherences.
6. The effect of the suicide rate on population decline.
7. Comparisons to other nations, particularly the US.
8. Administrative-level implementations.

Steps Taken to Address Problems

1. Distance units are in metric, and currency is in Korean Won.
2. Removed ‘gugun’, ‘dong’ from dataset 01.
    a. "index", "name", "buildDate", "min_sales", "max_sales", "avg_sales", "area", "floor", "pricePerArea" are in use.
3. Removed 'Urban_Population', 'Rural_Population', 'Population_Density' from dataset 02 KR.
    a. "Country", "Year", "Total_Population", "Life_Expectancy", "Birth_Rate", "Death_Rate", "Fertility_Rate", "Infant_Mortality_Rate", "Growth_Rate" are in use.
4. Removed 'Urban_Population', 'Rural_Population', 'Population_Density' from dataset 02 US.
    a. "Country", "Year", "Total_Population", "Life_Expectancy", "Birth_Rate", "Death_Rate", "Fertility_Rate", "Infant_Mortality_Rate", "Growth_Rate" are in use.
5. Removed 'Urban_Population', 'Rural_Population', 'Population_Density' from dataset 02 NG.
    a. "Country", "Year", "Total_Population", "Life_Expectancy", "Birth_Rate", "Death_Rate", "Fertility_Rate", "Infant_Mortality_Rate", "Growth_Rate" are in use.
6. Removed 'lat', 'lng' from dataset 03.
    a. "Id", "households", "buildDate", "score", "square_meters", "평", "min_sales", "max_sales", "avg_sales" are in use.

7. Removed ‘Region’ from dataset 04.
    a. "Date", "Birth", "Birth_rate", "Death", "Death_rate", “Divorce", "Divorce_rate", "Marriage", "Marriage_rate", "Natural_growth", "Natural_growth_rate" are in use.
8. Removed ‘sex’ from dataset 05 KR.
    a. "Country", "year", "sex", "age", "suicides_no" "population" are in use.
        i. "Country", "year", "age", "suicides_no", "population" are in use.
9. Removed ‘sex’ from dataset 05 US.
    a. "Country", "year", "sex", "age", "suicides_no" "population" are in use.
10. Removed 'wave', 'region', 'family_member', 'gender', 'religion', 'company_size', 'reason_none_worker' from dataset 06.
    a. "Id", "year", "income", "year_born", "education_level", "marriage", "occupation" are in use.
11. Removed dimensions of similar datasets accordingly.
12. Romanization of data in Korean language is not necessary as none appears in any of the visualizations.
13. Used means and linear models on multiple variables to determine significance.
14. Used plots for visualization to determine which variables have potential relationships
15. Combined data sets 01, 03
16. Combined data subsets KR, US

Analysis

1. Linear model of birth rate:
    a. Coefficients: (Intercept)
    b. Birth_Rate 2030.870
    c. -1.998
2. The mean birth rate of the ROK is 5.94
    a. Birth rate data spans from 2000 to 2020 displaying the rate of each month.
    b. The mean birth rate of the US is 11.34 nearly doubling that of the ROK.
    c. The mean birth rate of Niger is 45.84 more than quadrupling that of the US.
3. The mean divorce rate of the ROK is 2.351391 which is significantly lower than average and has declined in years, thus, is negatively correlated with birth rate. Based on this data, I can conclude that the divorce rate does not have a significant effect on birth rate.
4. Housing prices have steadily increased and have an average 82.18801 square meters with an average of 46451.58 units sold.Was unable to extract price trends over time based on current data.
5. Roughly two thirds of the population lives outside of the Seoul metropolitan area. Areas outside of Seoul are relevant in this study.
6. The mean of suicides is 703.58 per year. The mean of the total population is 51,585,058 the number of suicides is less than 0.00001 percent of the total population. This concludes that suicide rate is not relevant in this study.
7. Data is insufficient for issues such as social norms, trends, and intolerances.
8. Data is insufficient in determining the effects of the cost of living.

9. Property scores gradually increase from 1970 to 2020 with the max score being 5.0. Scores rose from 3.4 in the first quartile to 4.3 in the third quartile. Average sales also increased over the same period.
10. When making a direct visual comparison of the population trends of the ROK and US, the ROK somewhat levels off then decreases while the US steadily increases.

Implications

Results in the present and projections are daunting to say the least. The trajectory of the birth rate and population of the Republic of Korea are in significant decline when compared to the United States and even more so when compared to nations with greater increases than average. It is likely that social trends and the economic states of individuals are hugely impacting this; however, further research with additional data is required to confirm this. Variables such as housing costs, birth rates, marriage rates, and natural growth rates correlate negatively or positively all implying that the overall population is in a downward trajectory. As mentioned above, the projections look daunting, but I think it can be alleviated through stricter regulations in the housing market. I believe the main culprit to be rising real estate prices, thus I would undergo extensive research using more in-depth data sets regarding housing prices, cost of living, income, and government investments and funding in all facets of society.

Limitations

Despite using several datasets and multiple variables, I was severely limited due to the lack of relevant data. The main problem was continuous variables not spanning a sufficient time period. Variables such as birth rate, death rate, population, etc. only went back as far as 2000 generally. In order to conduct a more meaningful analysis, I would likely need data for these variables that go back at least 70 years. The reason for this is due to Korea's short modern history. The nation made its biggest economic and social transformation in the 1960s and 70s. Additional variables would also be needed in order to conduct further research. Variables relating to living costs over time, GDP, politics, and more in-depth economic features should be taken into consideration. Inconsistencies and unorganized data need to be handled in order to create more comprehensive and understandable visualizations.

Packages

library('ggplot2') - used for visualizations such as scatter plots.

library('stringr') - used for editing characters within strings.

library('purrr') - used to work with functions and vectors.

library('ggstatsplot') - used to generate dynamic visualizations.

library('Metrics') - used for regression and classification.

library('coefplot') - used to plot coefficients and standard errors.

library('dplyr') - used for managing and manipulating data sets.

library('tidyr') - used for managing and manipulating data sets.

Conclusion

The questions addressed were supported by data that have been used to validate the demographic problems that the Republic of Korea is already facing. As mentioned above, the lack of data renders the outcomes inconclusive and would require data covering a sufficient time period to lend any meaningful insights. Due to the complexity of the problem, other types of data such as demographics of the DPRK (North Korea) should also be considered. If I were to conduct this research more extensively, I would begin by adding relevant datasets for further analysis. Overall, insights have supported the already well-understood demographical issues that the ROK is currently facing.

GIT link: https://github.com/rlawnsdnjs706/DSC530_term.git