



MONASH
University

MONASH
BUSINESS
SCHOOL

Department of
Econometrics &
Business Statistics

☎ (03) 9905 2478
✉ BusEco-Econometrics@monash.edu

ABN: 12 377 614 012

Gender birth rate trend analysis in the United States

Arindam Baruah

abar0090@student.monash.edu (32779267)

Report for
ETC 5242 Task 1

1 August 2023



Contents

1	Introduction	2
2	Research Questions	2
3	Analysis	3
3.1	Query 1	3
3.2	Query 2	6
3.3	Query 3	6

1 Introduction

The current study delineates a detailed analysis of the birth counts for boys and girls born in the country of The United States of America. The study is based on two separate datasets, the first of which contain gender wise birth data from the years of **1629 to 1710** and the second, containing the gender birth data from the years of **1940 to 2002**. Utilisation of statistical techniques and subsequent feature engineering of the current and the historical data would further provide key insights into the changes of gender birth rate in the aforementioned country of interest. While the studies of Mathews, Hamilton, et al. (2005) and Ritchie and Roser (2019) have uncovered change of birth rates as a result of various socio-economic causes such as demographic changes, **the current study would however be limited to exploring the key changes in the number of births for girls and boys using statistical methods and feature engineering.**

It is critical to acknowledge that while the study by Pryzgoda and Chrisler (2000) defines the terminologies of “Sex ratio” and “Gender ratio” as two different statistics, however for the sake of simplicity, these terms may be used interchangeably in the context of the current study.

2 Research Questions

In this section of the study, the following key research questions pertaining to the new births in The United States will be formulated and further analysed in section 3 :

1. **How do the datasets belonging to the periods of 1629-1710 and 1940-2002 differ from one another ?**

2. How has the birth rate for girls changed over the period of 1940-2002 ?
3. Are there any similarities in the birth rate of girls between the time periods of 1629-1710 and 1940-2002 ?
4. Would creating new statistical features allow us to gain better insights into the data ?
5. Are boys born in greater proportion to girls during the period of 1940-2002 ?

3 Analysis

The current section will provide a step by step analysis for each of the formulated research questions of section 2.

3.1 Query 1

A basic understanding on the size of the data and the number of entries for the datasets belonging to each of the two time periods will be explored in this section.

```
head(df_present)
```

```
## # A tibble: 6 x 3
##   year    boys  girls
##   <dbl>  <dbl>  <dbl>
## 1  1940 1211684 1148715
## 2  1941 1289734 1223693
## 3  1942 1444365 1364631
## 4  1943 1508959 1427901
## 5  1944 1435301 1359499
## 6  1945 1404587 1330869
```

The above code-chunk and its output provides us with the glimpse of the dataset. The dataset contains three features. These features are explained as below:

- **Year** : The year pertaining to the count of new births in the United States.
- **boys** : The number of births classified as “boys” for the corresponding year.
- **girls** : The number of births classified as “girls” for the corresponding year.

Let us observe how do the values of the number of births between the periods of 1629-1710 and 1940-2002 vary from one another.

```
summary(df_present)
```

```
##      year      boys      girls
## Min.   :1940 Min.   :1211684 Min.   :1148715
## 1st Qu.:1956 1st Qu.:1799857 1st Qu.:1711404
## Median :1971 Median :1924868 Median :1831679
## Mean   :1971 Mean   :1885600 Mean   :1793915
## 3rd Qu.:1986 3rd Qu.:2058524 3rd Qu.:1965538
## Max.   :2002 Max.   :2186274 Max.   :2082052
```

```
summary(df_arb)
```

```
##      year      boys      girls
## Min.   :1629 Min.   :2890 Min.   :2722
## 1st Qu.:1649 1st Qu.:4759 1st Qu.:4457
## Median :1670 Median :6073 Median :5718
## Mean   :1670 Mean   :5907 Mean   :5535
## 3rd Qu.:1690 3rd Qu.:7576 3rd Qu.:7150
## Max.   :1710 Max.   :8426 Max.   :7779
```

As we can clearly observe, the **magnitude of the boys and girls born during the period of 1940-2002 are much larger than that for the period of 1629-1710**. This is expected as a result of the global rise of population owing to factors such as better infrastructure, better lifestyle, better socio-economic factors and improvement in medical sciences.

Let us try to visualise the same through figure 1.

```
options(scipen = 999) # To remove scientific notation

df_arb_long <-
  pivot_longer(
    df_arb,
    names_to = "gender",
    values_to = "born",
    cols = c(boys, girls)
  )
p11 <-
  ggplot(data = df_arb_long, aes(x = year, y = born, fill = gender)) +
  geom_area(color = 'black') + theme_classic() +
  ggtitle("Birth statistics between 1629-1710 \n in The US") +
  theme(plot.title = element_text(hjust = 0.5), aspect.ratio = 0.5) +
  labs(fill = "Gender", x = "Year", y = "Number of births")

df_present_long <-
  pivot_longer(
```

```
df_present,
names_to = "gender",
values_to = "born",
cols = c(boys, girls)
)
p12 <-
ggplot(data = df_present_long, aes(x = year, y = born, fill = gender)) +
geom_area(color = 'black') +
theme_classic() +
ggtitle("Birth statistics between 1940-2002 \n in The US") +
theme(plot.title = element_text(hjust = 0.5),aspect.ratio = 0.5) +
labs(fill = "Gender", x = "Year", y = "Number of births")

plot_grid(p11, p12, labels = "AUTO", ncol = 1)
```

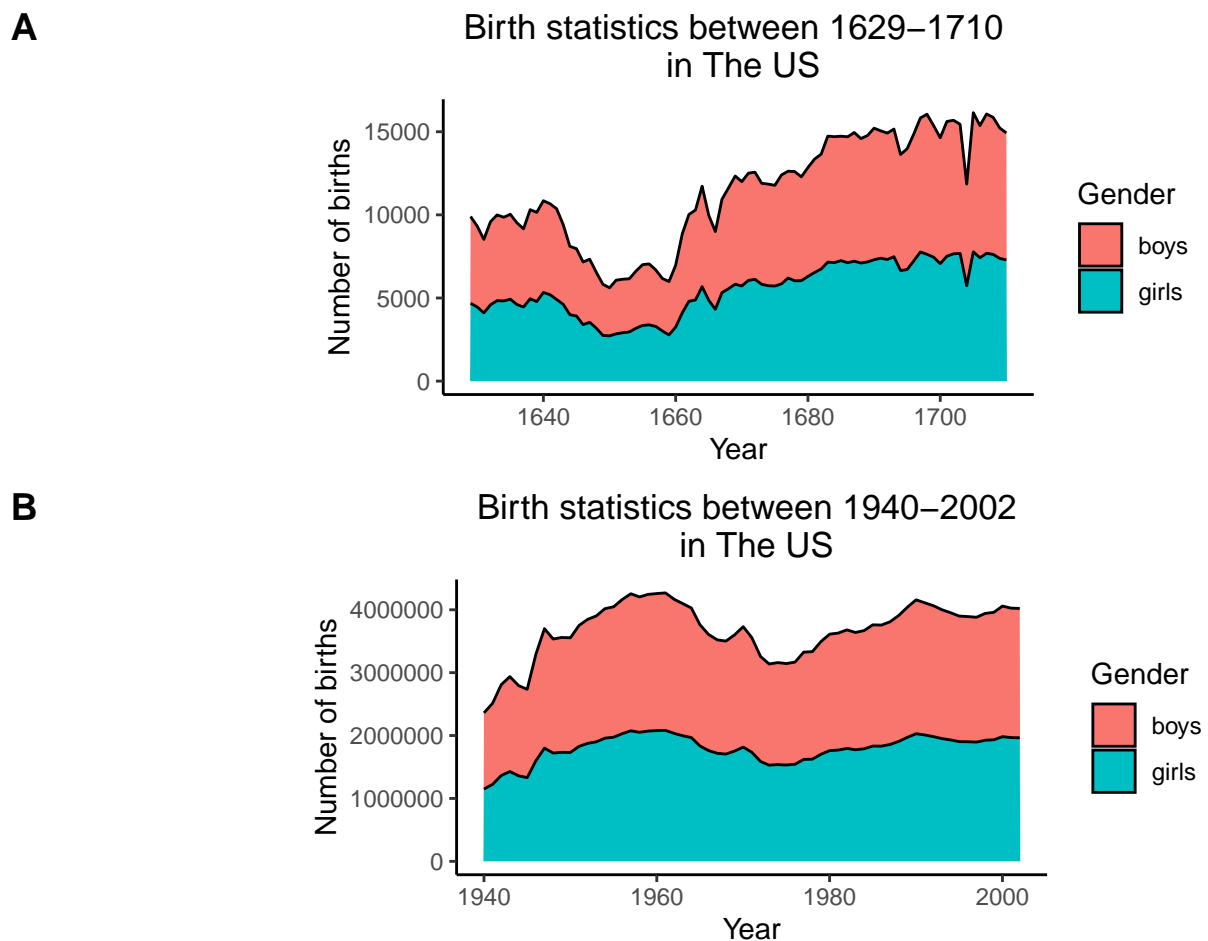


Figure 1: *Number of births in The United States of America*

As we can observe through figure 1, the magnitude of births are observed to be **significantly higher** during the period between 1940-2002 when compared to the period between 1620-1710.

3.2 Query 2

In this section, the number of births classified as “girls” will be focused on to gain further insights through a scatter plot as visualised through figure 2.

```
p13 <- ggplot(data = df_present, aes(x = year, y = girls)) +  
  geom_point(shape = 2, color = 'red') + theme_classic() +  
  ggtitle("Number of girls born during 1940-2002 \n in the US") +  
  theme(plot.title = element_text(hjust = 0.5), aspect.ratio = 0.5) + labs(x = "Year",  
                                                                           y = "Number of births")  
  
p14 <- ggplot(data = df_arb, aes(x = year, y = girls)) +  
  geom_point(shape = 2, color = 'blue') + theme_classic() +  
  ggtitle("Number of girls born during 1629-1710 \n in the US") +  
  theme(plot.title = element_text(hjust = 0.5), aspect.ratio = 0.5) + labs(x = "Year",  
                                                                           y = "Number of births")  
  
plot_grid(p13,p14,ncol =1)
```

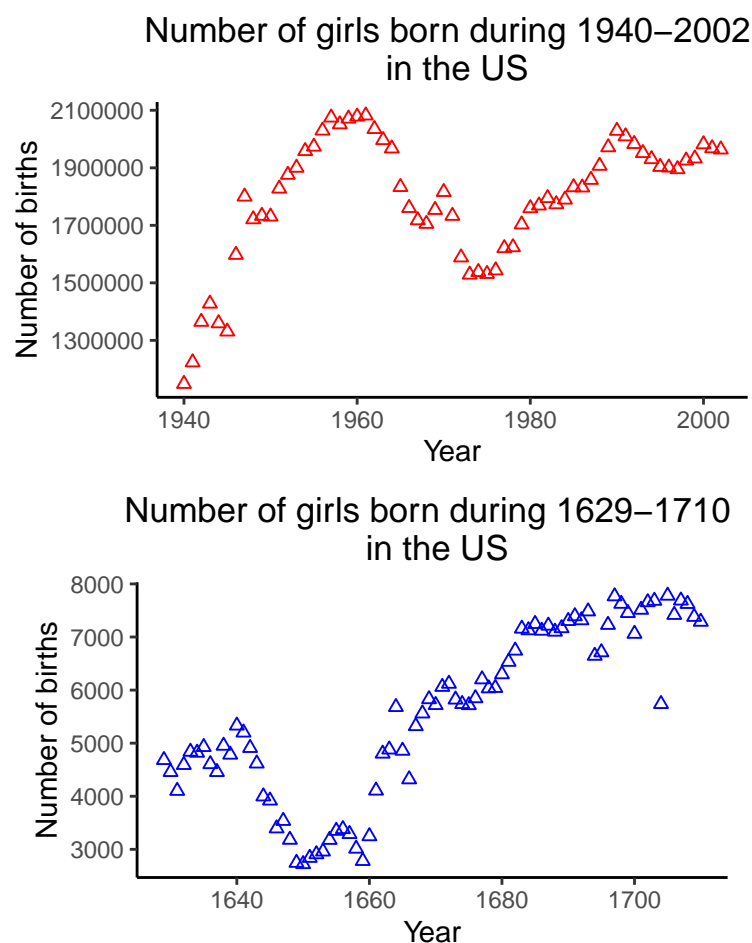


Figure 2: Girls born during each period in the US

3.3 Query 3

Based on the analysis of figures 2 and 3, we can observe the following key points:

1. The magnitude of girls born in the period between 1940-2002 is **significantly higher** than in the period between 1629-1710.

2. During the period between 1940-1960, there was a **consistent increase in the number of girls born in the country**.
3. The period between 1960-1975 observed a **sudden drop in the number of girls born**.
4. However, the number of girls born were **back on the rise** after the year 1975.
5. Between the period of 1990-2000, the **number of girls born were observed to stagnate at approximately 2000000 births a year**.

```
pl5 <- ggplot(data = df_present, aes(x = year, y = girls)) +  
  geom_point(shape = 2, color = 'red') + theme_classic() +  
  ggtitle("Number of girls born during 1940-2002 \n in the US") +  
  labs(x = "Year", y = "Number of births") + theme(plot.title = element_text(hjust = 0.5),  
    aspect.ratio = 0.7) +  
  
  annotate(  
    "segment",  
    x = 1945,  
    y = 1900000,  
    xend = 1950 ,  
    yend = 1800000,  
    arrow = arrow(type = "closed",  
      length = unit(0.02, "npc"))  
  ) +  
  annotate(  
    "text",  
    x = 1946,  
    y = 1980000,  
    colour = "black",  
    label = 'Consistent rise in number of girls \n born between 1940-1960',  
    size = unit(2.5, "pt")  
  ) +  
  annotate(  
    "segment",  
    x = 1965,  
    y = 1300000,  
    xend = 1973 ,  
    yend = 1500000,  
    arrow = arrow(type = "closed",  
      length = unit(0.02, "npc"))  
  ) +  
  annotate(  
    "text",  
    x = 1965,  
    y = 1250000,  
    colour = "blue",  
    label = 'Dip in number of girls born \n between 1960-1975',  
    size = unit(2.5, "pt")  
  ) +  
  annotate(  
    "segment",  
    x = 1983,  
    y = 1500000,  
    xend = 1985 ,  
    yend = 1700000,  
    arrow = arrow(type = "closed",  
      length = unit(0.02, "npc"))  
  ) +  
  annotate(  
    "text",  
    x = 1985,  
    y = 1450000,  
    colour = "darkgreen",  
    label = 'Girl child births start \n increasing again',  
    size = unit(2.5, "pt")  
  )
```

```
) +
  annotate(
    "segment",
    x = 1995,
    y = 1700000,
    xend = 1995,
    yend = 1850000,
    arrow = arrow(type = "closed",
                  length = unit(0.02, "npc"))
  ) +
  annotate(
    "text",
    x = 1995,
    y = 1600000,
    colour = "darkred",
    label = 'Stagnation in the number \n of girls born after 1990',
    size = unit(2.5, "pt")
  )
)
p15
```

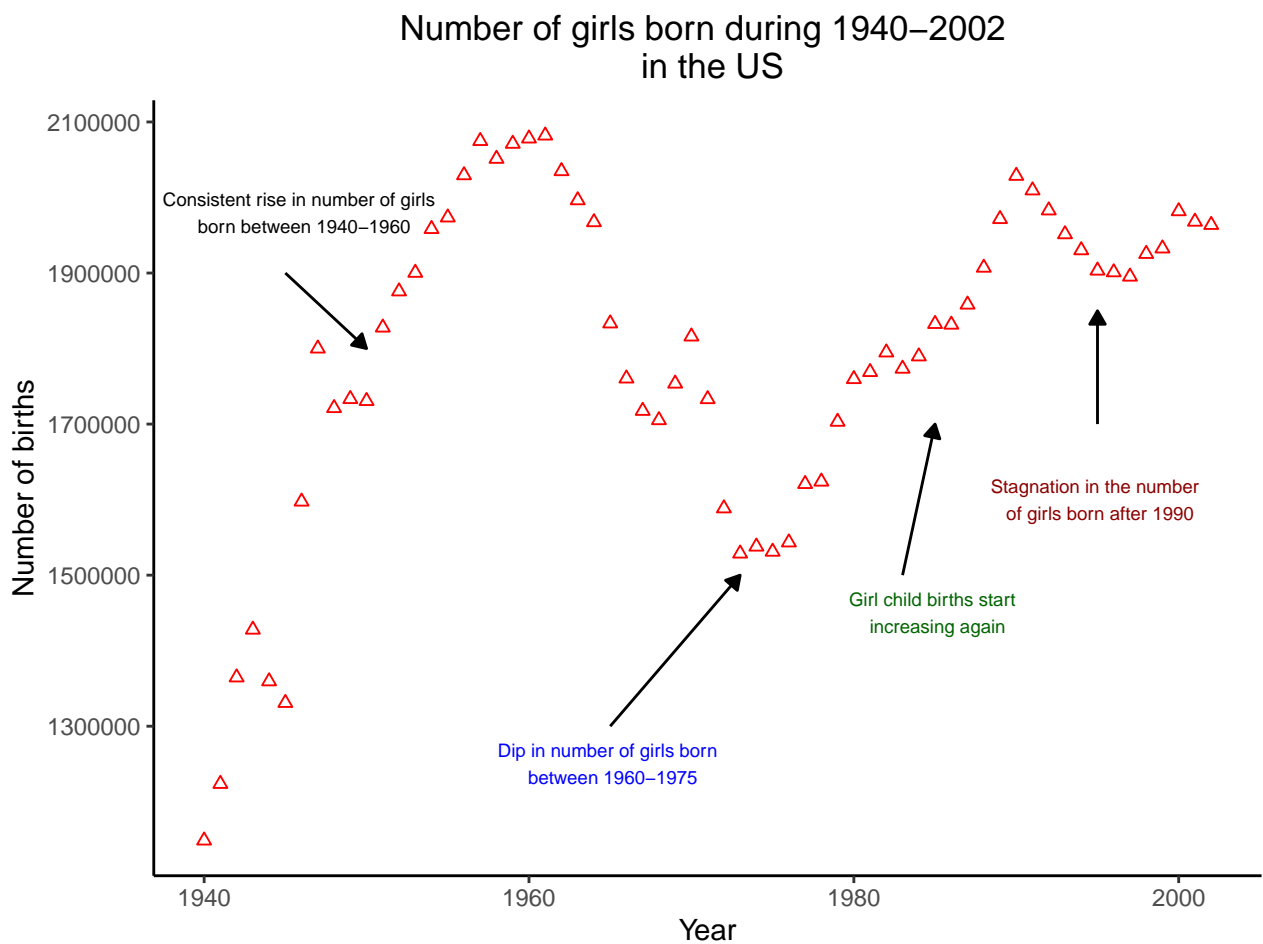


Figure 3: Trend analysis of girls child births in The United States during 1940-2002

References

Mathews, T, BE Hamilton, et al. (2005). Trend analysis of the sex ratio at birth in the United States. *National vital statistics reports* **53**(20), 1–17.

Pryzgoda, J and JC Chrisler (2000). Definitions of gender and sex: The subtleties of meaning. *Sex roles* **43**, 553–569.

Ritchie, H and M Roser (2019). Gender Ratio. *Our World in Data*. <https://ourworldindata.org/gender-ratio>.