Gender Equality in Primary and Secondary Education

Project for STAT 613

Xiaonan Zhang

April 24, 2018

When we talk about gender equality, we always pay attention to college admissions, employee ratios, equal pay for equal work, or glass ceilings. These are important, of course. It's just one day I asked myself: if a country owns a low ratio on its primary and secondary education, how could it be possible to have a high ratio on college education? Basically the ratio of primary and secondary education influences the ratio of college and employees, then leadership and political participation as well. Hence, learning about the current circumstance in all countries could really help us make a better decision on what could we do to improve gender equality. This project is about to identify different countries, divide them into natural groups, and help decision makers understand what's going on better.

All the data come from world bank. You could only download the data for one country one time from world bank, so I download the integrated version from gapminder. I put links within the rmarkdown file, so feel free to try it out.

To be honest, the initial plan is to analyze gender ratio in primary and secondary education. Take a look at the ratio of girls to boys (%) for all 197 countries and their expenditure on students. If a family cannot afford all children to join a school, the subsidy may give them more incentive to get their children to school. However, the expenditure data is not available for all countries. I tried to find expenditure per student in primary, secondary, and tertiary education. All these data are available on world bank, but all of these data are not enough for analysis. Shown below by sum function.

I could find more variables to explain gender inequality, but I'm not comfortable doing that. In my understanding, statistics should provide methodology but based on the specific field. I don't want to use either inner_join or picking other variables. What else variables could it be possible? Like income per family or GDP per capital? Maybe it's just about social culture and nothing to do with government expenditure on education? Or even the expenditure may enlarge the ratio because they encourage more boys rather than girls to go to school.

For a data science project, I don't want to focus on what reasons lead to gender inequality or what policy could help improve equality. I prefer to focus on one dataset and do more analysis based on it.

Let's get started!

Install the packages first.

```
library(readxl)
library(tidyverse)
library(countrycode)
library(modelr)
library(gridExtra)
library(broom)
```

Download the data online and tidy them into one data frame. Feel free to rerun this code. Don't worry. All files downloaded here will be automatically deleted in the end.

```
if(!file.exists("./edustat")){dir.create("./edustat")}
fileUrls <- c(
  "https://docs.google.com/spreadsheet/pub?key=pyj6tScZqmEcWM3hb0x-BZA&output=xlsx",
  "https://docs.google.com/spreadsheet/pub?key=0AkBd6lyS3EmpdE8xR0dUWDI4ME02SjQ5bi1NYnFHN0E&output=xlsx
```

```
  "https://docs.google.com/spreadsheet/pub?key=0AkBd6lyS3EmpdFJTUEVleTM0cE5jTnlTMk41ajBGclE&output=xlsx
  "https://docs.google.com/spreadsheet/pub?key=0AkBd6lyS3EmpdDBuUVIzbWwtaU5helpJVG5BMmxyX1E&output=xlsx
  "https://docs.google.com/spreadsheet/pub?key=0AkBd6lyS3EmpdDJxdlN6cEtYMjMxdC1XdGdKOXR2bkE&output=xlsx
  )
varNames <- c("gb_edu","gb_1524","exp_pri","exp_sec","exp_ter")
get_cleaned <- function(url_in,var_name){
  download.file(url_in,destfile = "./edustat/tmp.xlsx",mode = "wb")
  output <- readxl::read_excel("./edustat/tmp.xlsx")
  names(output)[1] <- "country"
  output <- gather(output,key = year ,value = !!var_name, -1, na.rm = TRUE) %>%
    arrange(country)
  output <- mutate(output,year=as.integer(year))
}

all_data <- map2(fileUrls, varNames, get_cleaned)
#head(all_data)
#test whether all functions above works

df <- reduce(all_data,left_join) %>% as.data.frame()
df$continent <-
  as.factor(countrycode(
    sourcevar = df[,"country"],origin = "country.name", destination = "continent"))
```

Add continent information into the dataset. I want to do some analysis based on different continents.

The ratio dataset contains 4657 observations, but the expenditure is only about 750 data, most years of data of most of the countries are unavailable.

```
sum(!is.na( df$gb_edu ))
```

```
## [1] 4657
```

```
sum(!is.na(df$gb_1524))
```

```
## [1] 332
```

```
sum(!is.na(df$exp_pri))
```

```
## [1] 789
```

```
sum(!is.na(df$exp_sec))
```

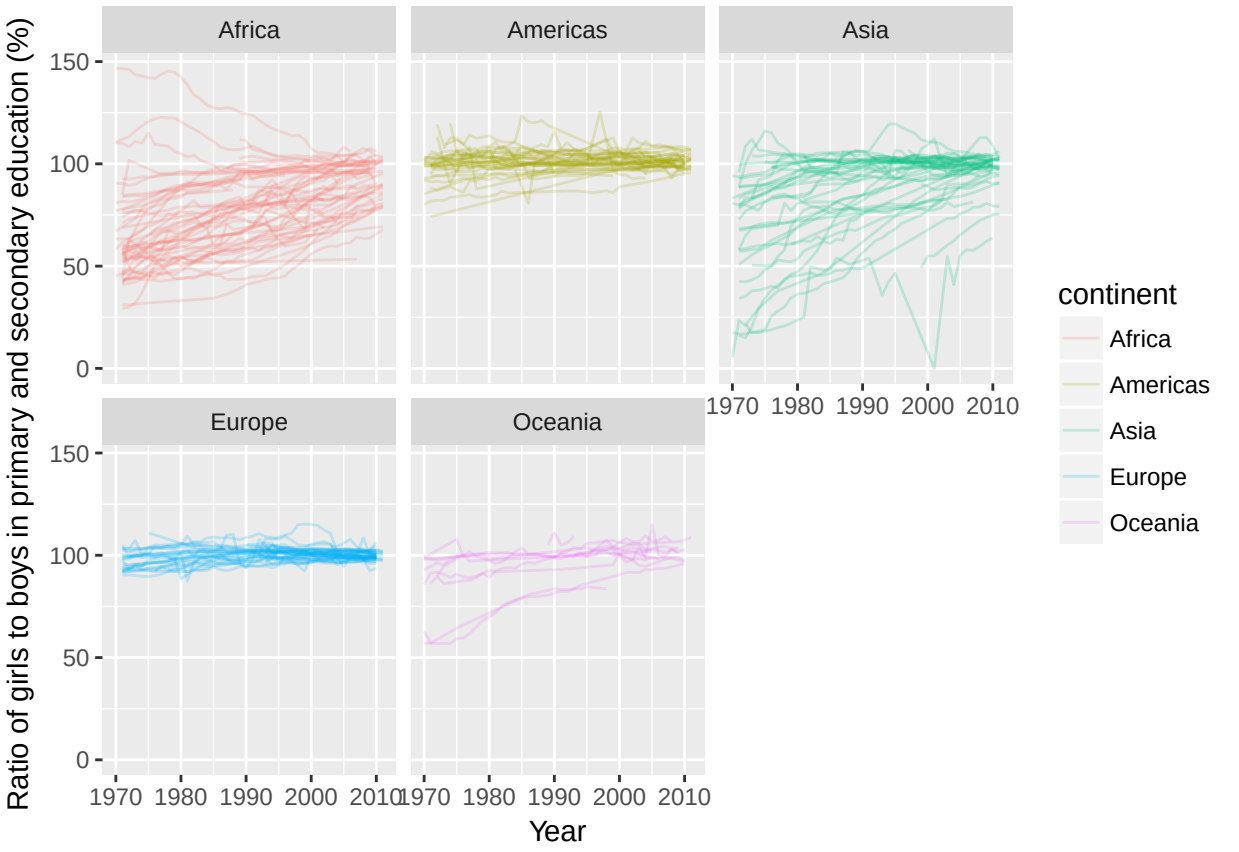
```
## [1] 760
```

```
sum(!is.na(df$exp_ter))
```

```
## [1] 715
```

Then let's focus on "ratio of girls to boys in primary and secondary education" dataset and determine which countries alert us.

```
df %>%
  ggplot(aes(x=year,y= gb_edu, group=country,color=continent))+
  geom_line(alpha=0.2)+
  facet_wrap("continent",scales = "fixed")+
  ylab("Ratio of girls to boys in primary and secondary education (%)")+
  xlab("Year")
```

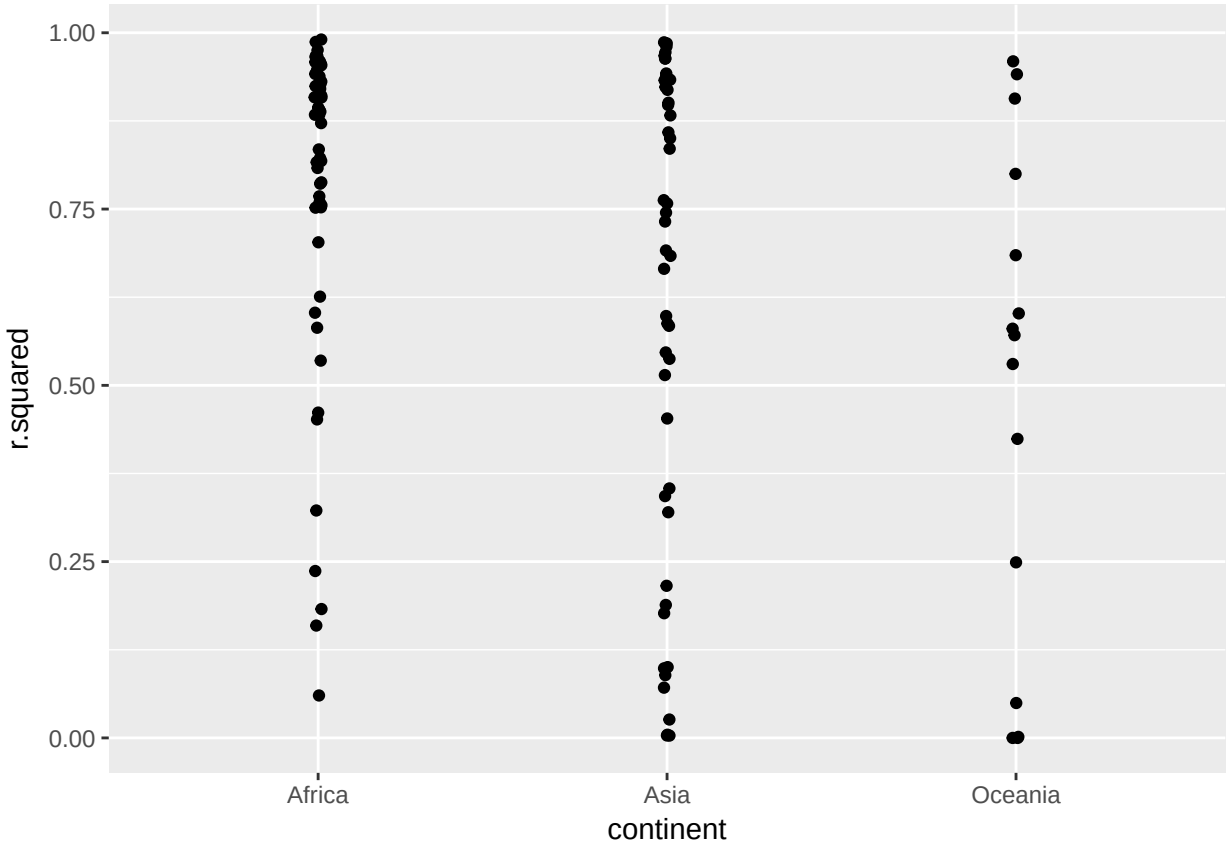


With no surprises, Europe always has a good performance. There are lots of stuff we need to worry about Europe, but not gender equality in primary and secondary education. The most interesting pattern shows in Africa, Asia, and Oceania. We want to distinguish the bad performance countries from this messy plot.

Notice that most countries show an increasing trend over time. Try to build a model over time for each country. We have 197 countries in our dataset, so we definitely do not want to build the model one by one. Use nest() and map() here to go through all countries.

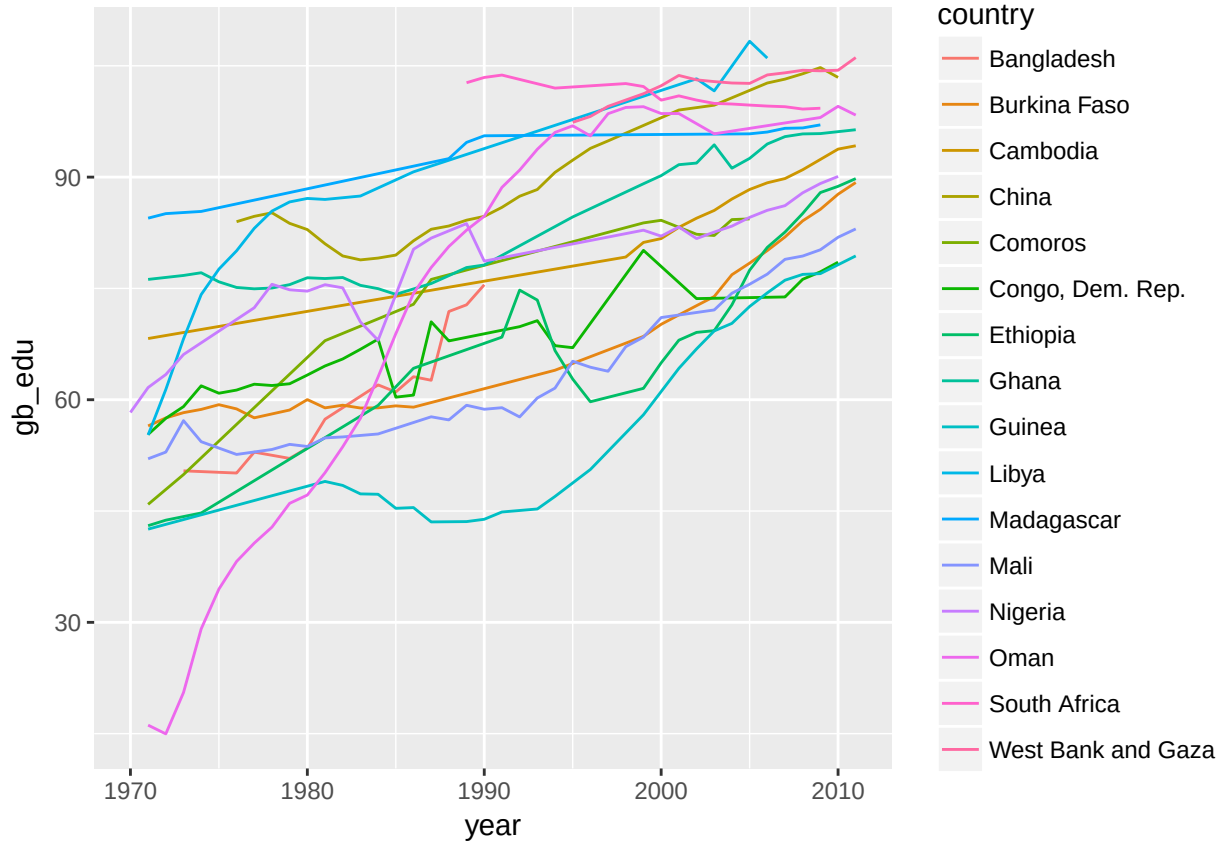
```
tidy_df <- df %>%
  group_by(country,continent)%>%
  nest()
tidy_df <- tidy_df[tidy_df$continent!="Europe" & tidy_df$continent!="Americas",]
mdl <- function(df){
  lm(gb_edu~year,data=df)
}
#all_model <- map(tidy_df$data,mdl)
tidy_df <- tidy_df %>%
  mutate(model=map(data,mdl))

mdl_fit <- tidy_df %>%
  mutate(mdlfit=map(model,glance)) %>%
  unnest(mdlfit,.drop = TRUE)%>%
  arrange(r.squared)
mdl_fit %>%
  ggplot(aes(x=continent,y=r.squared)) +
  geom_jitter(width = 0.01)
```



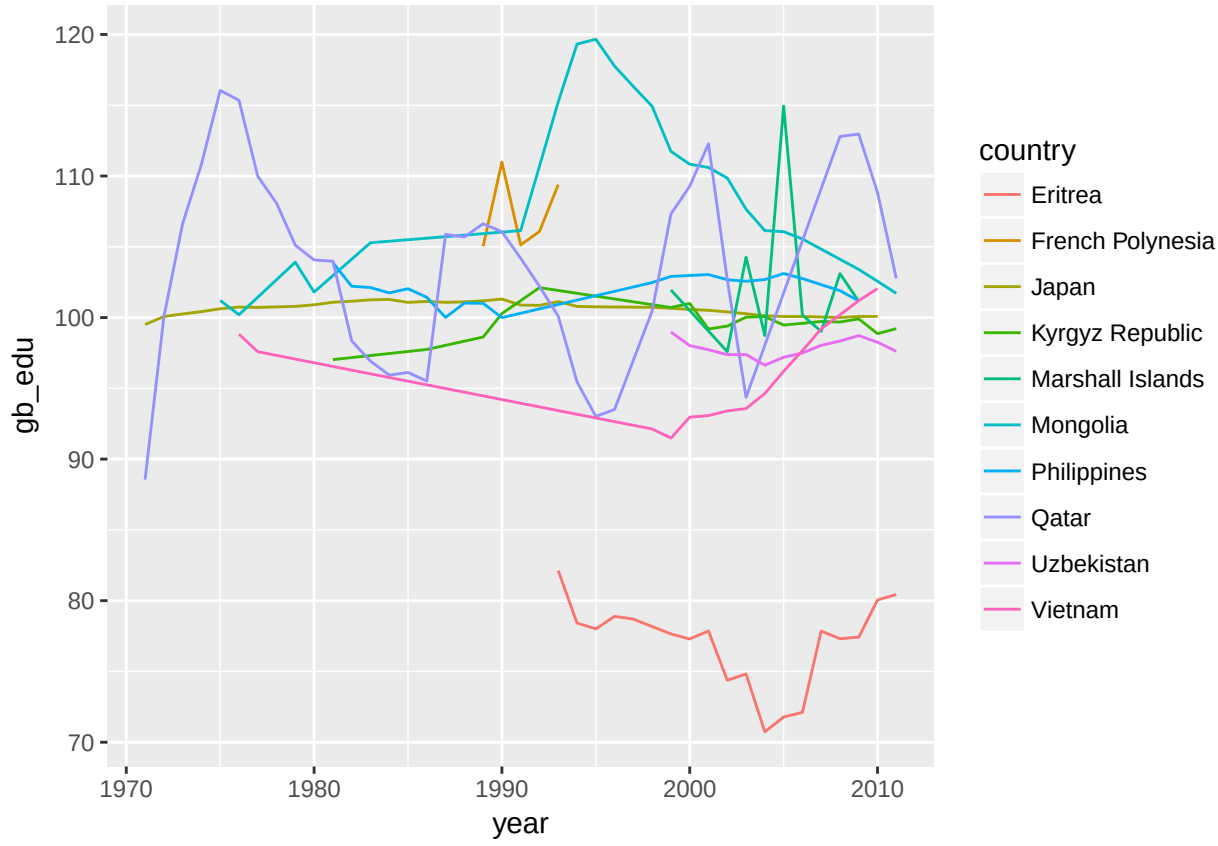
In our case, the higher r.square means the ratio in this country increases over time. Do we need to worry about those counties? It depends. On one hand, they are not bad since they have an increasing trend. On the other hand, could we do something to improve the increasing speed and help them get to 1:1 sooner?

```
mdl_select <- filter(mdl_fit, r.squared > 0.8 & r.squared < 0.9)
df %>%
  semi_join(mdl_select,by="country") %>%
  ggplot(aes(year,gb_edu,color=country))+
  geom_line()
```



For countries where the ratio does not increase over time, we do need to take a look and find a way to improve their situation. Select small r.square to pick up these countries.

```
mdl_select <- filter(mdl_fit, r.squared > 0.0 & r.squared < 0.1)
df %>%
  semi_join(mdl_select,by="country") %>%
  ggplot(aes(year,gb_edu,color=country))+
  geom_line()
```



Interesting facts occur, the ratio of some countries is floating around 100%. What happened? We need to search their historical stories. Models help up find these countries, but not be able to tell us about the history.

Besides figuring out the trend over time, the current circumstance is also crucial for us to make any decisions. The question is what ratio should we use as a threshold for low-ratio countries? 0.3 or 0.5? I answer this question by 'cluster'. With the help of cluster, we don't need one fixed threshold. We could divide countries by their natural groups.

The cluster method and output are within the shiny app.

https://xiaoooonan.shinyapps.io/SecondPartOf613/