R Notebook : United Nations Life expectancy data analysis

Code

By Palash Khandekar

Hide

```
# Loading packages
library(dplyr)
library(tidyr)
library(ggplot2)
# Loading data
life_expectancy <- read.csv("data.csv")
# Taking a look at the first few rows
head(life_expectancy)
```

	Country.or.Area <fctr>	Subgroup <fctr>	Year <fctr>
1	Afghanistan	Female	2000-2005
2	Afghanistan	Female	1995-2000
3	Afghanistan	Female	1990-1995
4	Afghanistan	Female	1985-1990
5	Afghanistan	Male	2000-2005
6	Afghanistan	Male	1995-2000

6 rows | 1-4 of 7 columns

Life expectancy of men vs. women by country

Let’s manipulate the data to make our exploration easier. We will build the dataset for our first plot in which we will represent the average life expectancy of men and women across countries for the last period recorded in our data (2000-2005).

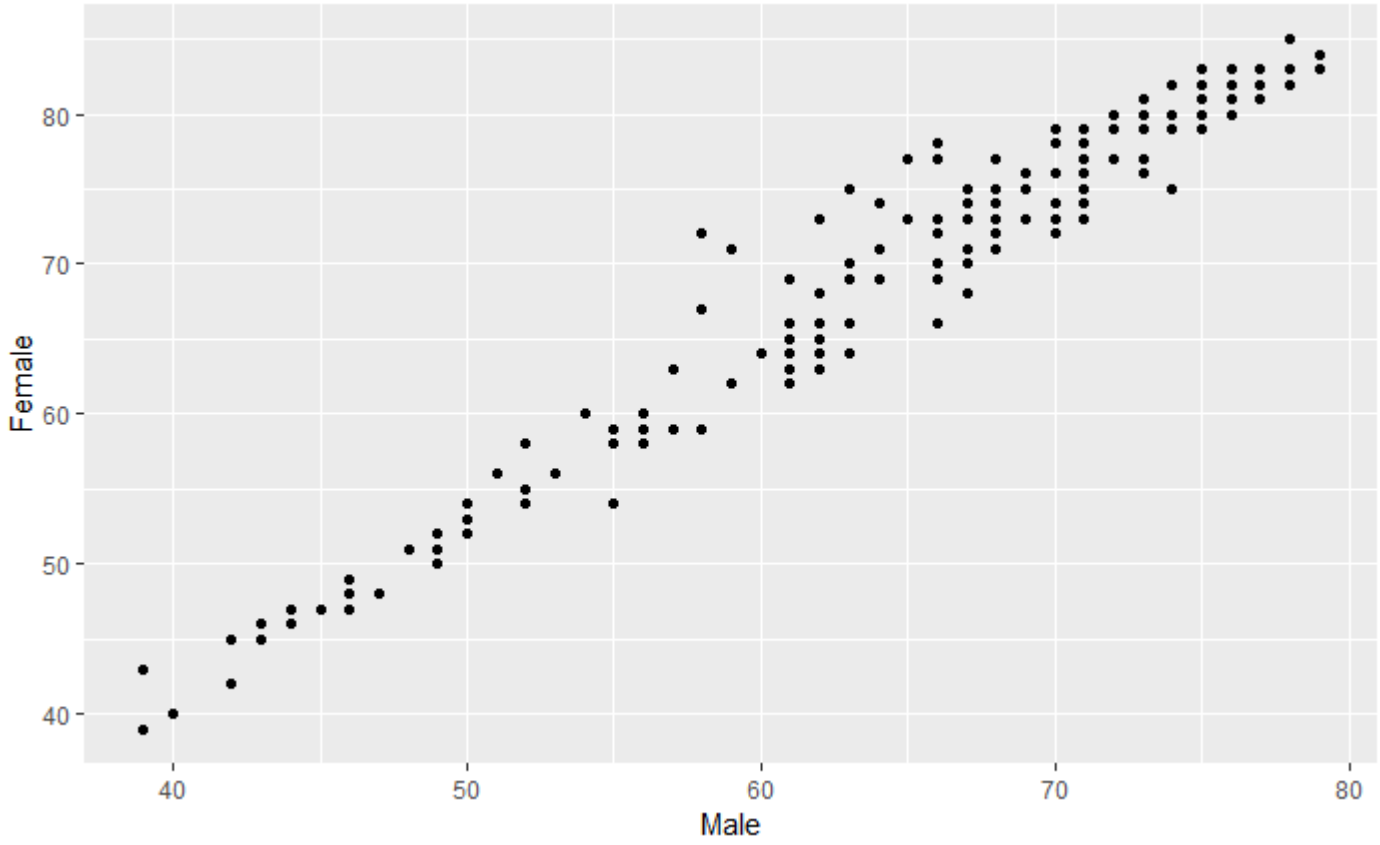
Hide

```
# Subsetting and reshaping the life expectancy data
subdata <- life_expectancy  %>%
    filter(Year == '2000-2005')%>%
    select(Country.or.Area, Subgroup, Value) %>%
    spread(Subgroup, Value)
# Taking a look at the first few rows
head(subdata)
```

	Country.or.Area <fctr>	Female <int>	Male <int>
1	Afghanistan	42	42
2	Albania	79	73
3	Algeria	72	70
4	Angola	43	39
5	Argentina	78	71
6	Armenia	75	68

6 rows

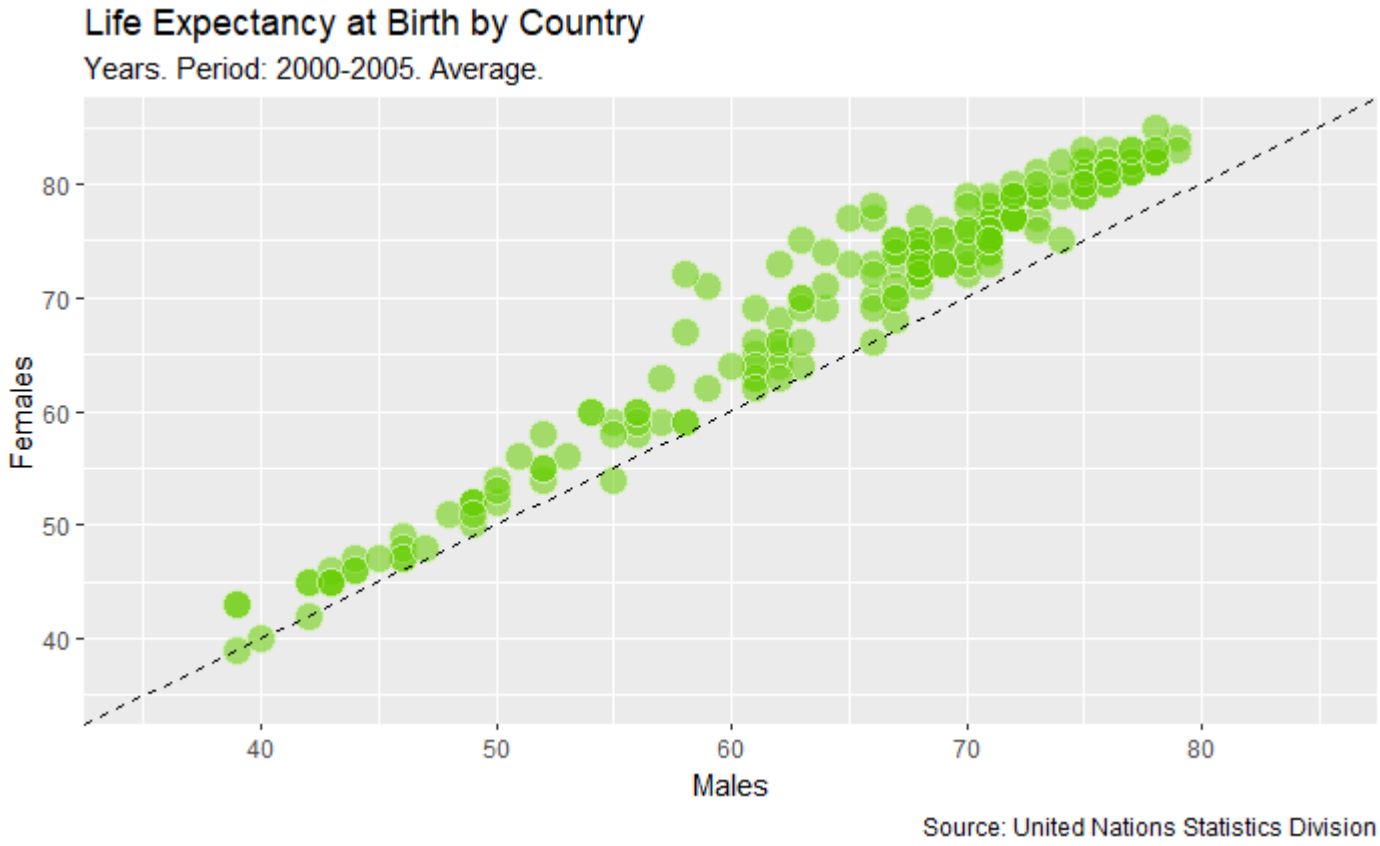
Let's create a scatter plot using `ggplot2` to represent life expectancy of males (on the x-axis) against females (on the y-axis). We will create a straightforward plot in this task, without many details. We will take care of these kinds of things shortly.



Reference lines I

Hide

```
# Adding an abline and changing the scale of axes of the previous plots
# Adding labels to previous plot
ggplot(subdata, aes(x=Male, y=Female))+
  geom_point(colour="white", fill="chartreuse3", shape=21, alpha=.55, size=5)+
  geom_abline(intercept = 0, slope = 1, linetype=2)+
  scale_x_continuous(limits=c(35,85))+
  scale_y_continuous(limits=c(35,85))+
  labs(title="Life Expectancy at Birth by Country",
       subtitle="Years. Period: 2000-2005. Average.",
       caption= "Source: United Nations Statistics Division",
       x="Males",
       y="Females")
```

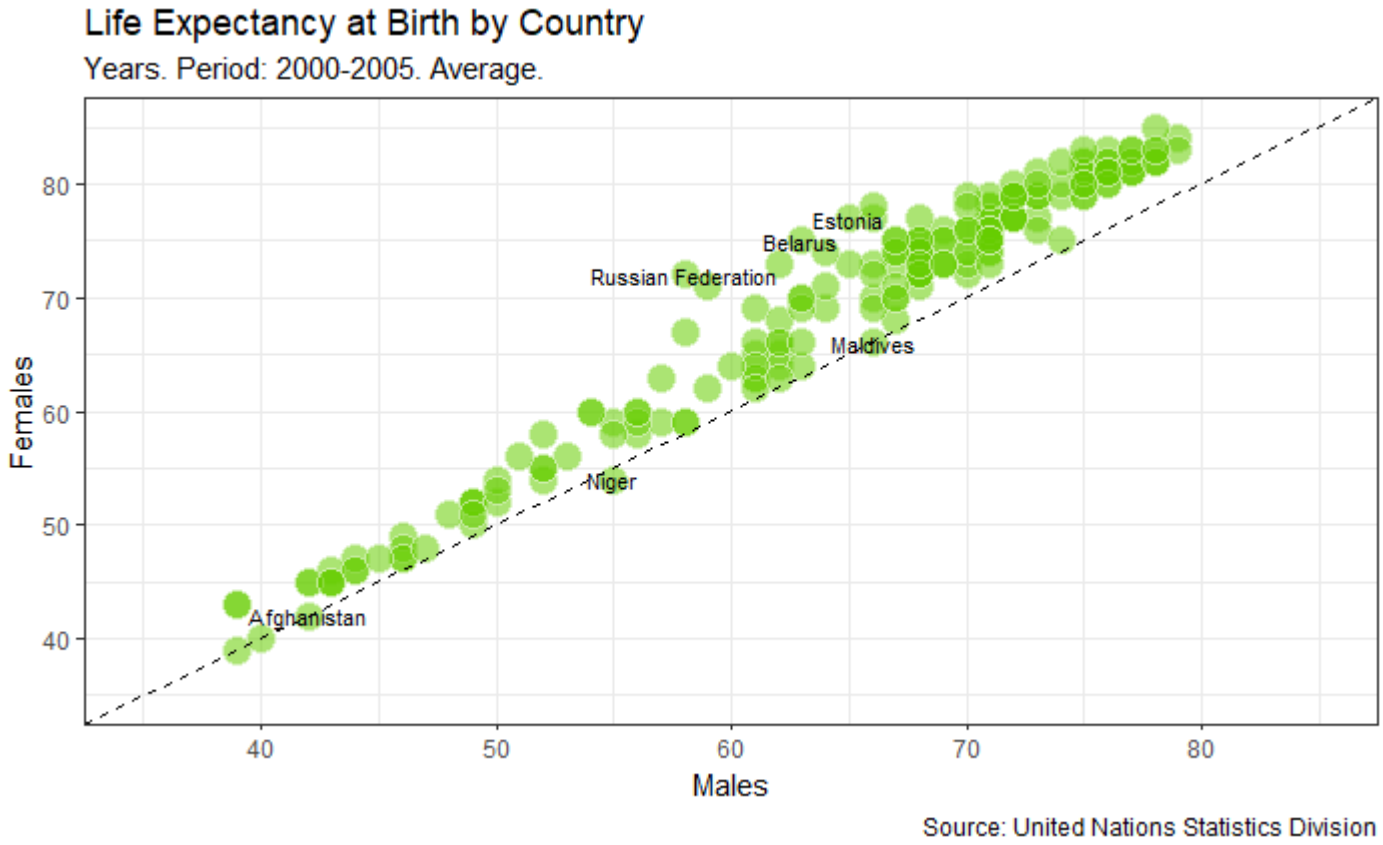


Highlighting remarkable countries I

Now, we will label some points of our plot with the name of its corresponding country. We want to draw attention to some special countries where the gap in life expectancy between men and women is significantly high. These will be the final touches on this first plot.

Hide

```
# Subseting data to obtain countries of interest
top_male <- subdata %>% arrange(Male-Female) %>% head(3)
top_female <- subdata %>% arrange(Female-Male) %>% head(3)
# Adding text to the previous plot to label countries of interest
ggplot(subdata, aes(x=Male, y=Female, label=Country.or.Area))+
  geom_point(colour="white", fill="chartreuse3", shape=21, alpha=.55, size=5)+
  geom_abline(intercept = 0, slope = 1, linetype=2)+
  scale_x_continuous(limits=c(35,85))+
  scale_y_continuous(limits=c(35,85))+
  labs(title="Life Expectancy at Birth by Country",
       subtitle="Years. Period: 2000-2005. Average.",
       caption="Source: United Nations Statistics Division",
       x="Males",
       y="Females")+
    geom_text(data=top_male, size=3)+
geom_text(data=top_female, size=3)+
theme_bw()
```



How has life expectancy by gender evolved?

Since our data contains historical information, let's see now how life expectancy has evolved in recent years. Our second plot will represent the difference between men and women across countries between two periods: 2000-2005 and 1985-1990.

Let's start building a dataset called `subdata2` for our second plot.

Hide

```
# Subsetting, mutating and reshaping the life expectancy data
subdata2 <- life_expectancy %>%
  filter(Year %in% c("1985-1990", "2000-2005")) %>%
  mutate(Sub_Year=paste(Subgroup, Year, sep="_")) %>%
  mutate(Sub_Year=gsub("-", "_", Sub_Year)) %>%
  select(-Subgroup, -Year) %>%
  spread(Sub_Year, Value)%>%
  mutate(diff_Female = Female_2000_2005 - Female_1985_1990) %>%
  mutate(diff_Male = Male_2000_2005 -  Male_1985_1990)

# Taking a look at the first few rows
head(subdata2)
```

	Country.or.Area <fctr>	Source <fctr>
1	Afghanistan	UNPD_World Population Prospects_2006 (International estimate)
2	Albania	UNPD_World Population Prospects_2006 (International estimate)
3	Algeria	UNPD_World Population Prospects_2006 (International estimate)
4	Angola	UNPD_World Population Prospects_2006 (International estimate)
5	Argentina	UNPD_World Population Prospects_2006 (International estimate)
6	Armenia	UNPD_World Population Prospects_2006 (International estimate)

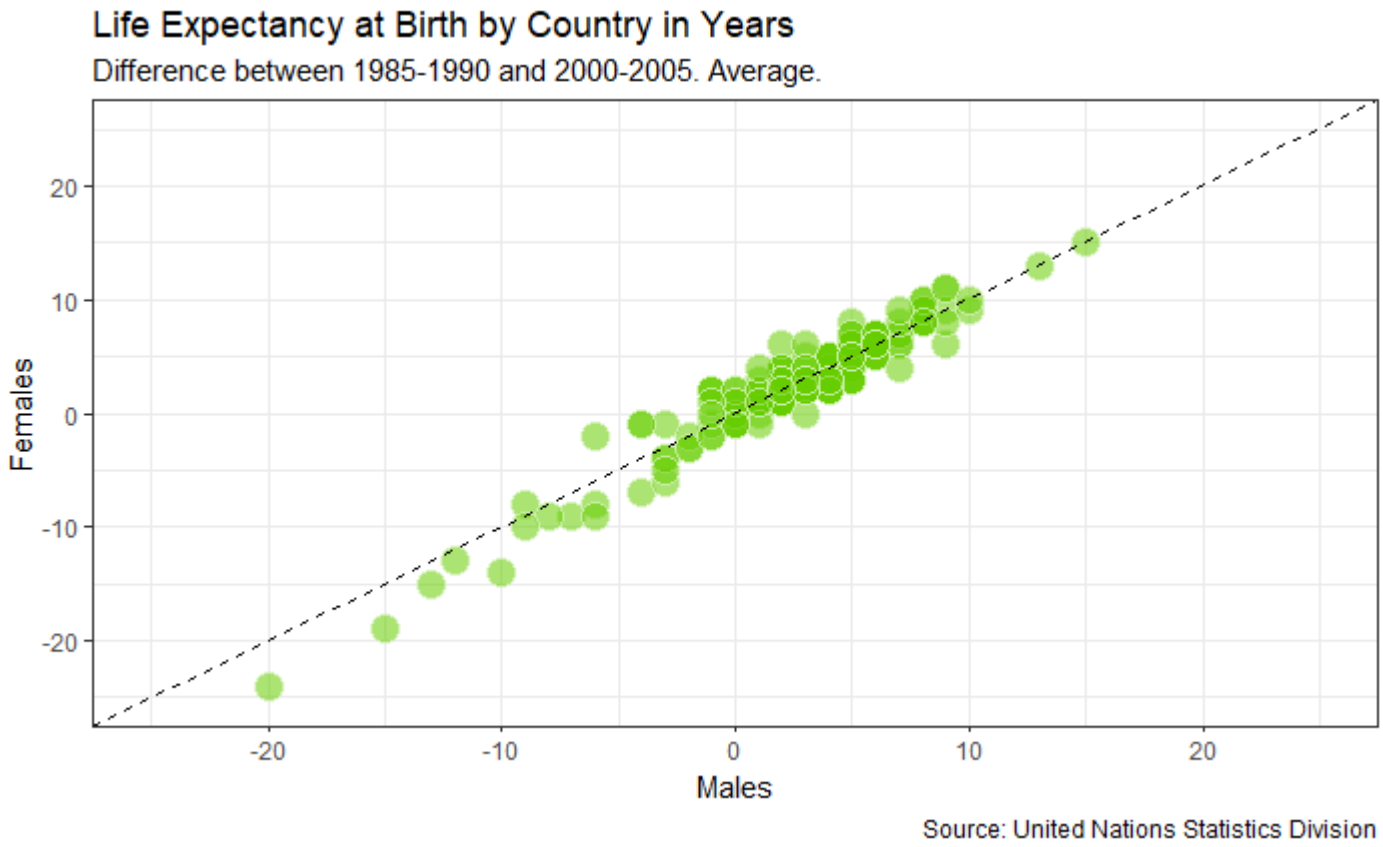
6 rows | 1-4 of 10 columns

Visualize

Now let's create our second plot in which we will represent average life expectancy differences between "1985-1990" and "2000-2005" for men and women.

Hide

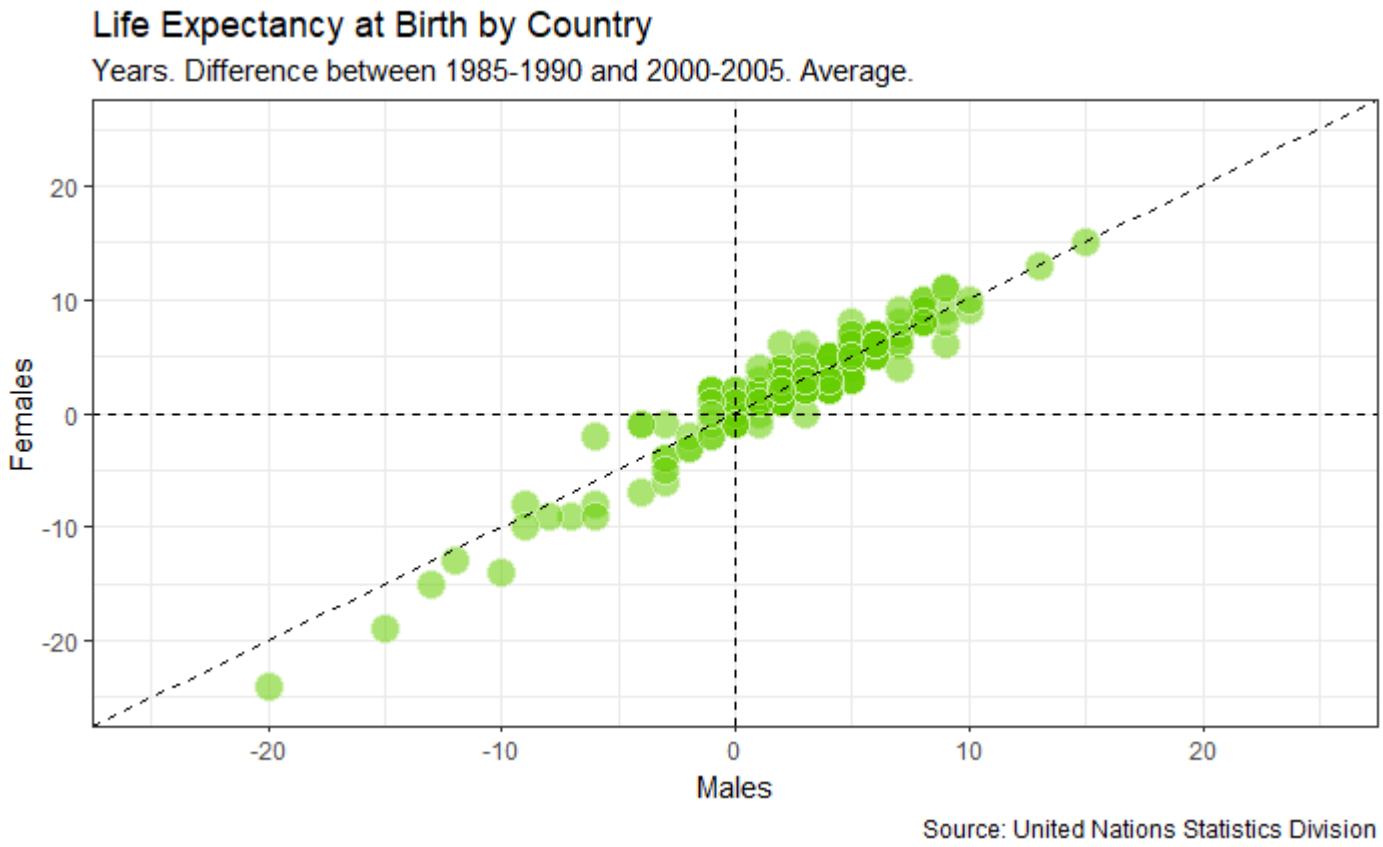
```
# Doing a nice first version of the plot with abline, scaling axis and adding labels
ggplot(subdata2, aes(x=diff_Male, y=diff_Female, label=Country.or.Area))+
  geom_point(colour="white", fill="chartreuse3", shape=21, alpha=.55, size=5)+
  geom_abline(intercept = 0, slope = 1, linetype=2)+
  scale_x_continuous(limits = c(-25,25))+
  scale_y_continuous(limits = c(-25,25))+
  labs(title="Life Expectancy at Birth by Country in Years",
       subtitle="Difference between 1985-1990 and 2000-2005. Average.",
       caption="Source: United Nations Statistics Division",
       x="Males",
       y="Females")+
theme_bw()
```



Reference lines II

Hide

```
# Adding an hline and vline to previous plots
ggplot(subdata2, aes(x=diff_Male, y=diff_Female, label=Country.or.Area))+
  geom_point(colour="white", fill="chartreuse3", shape=21, alpha=.55, size=5)+
  geom_abline(intercept = 0, slope = 1, linetype=2)+
  scale_x_continuous(limits=c(-25,25))+
  scale_y_continuous(limits=c(-25,25))+
  geom_vline(xintercept = 0, linetype = 2)+
  geom_hline(yintercept = 0, linetype = 2)+
  labs(title="Life Expectancy at Birth by Country",
       subtitle="Years. Difference between 1985-1990 and 2000-2005. Average.",
       caption="Source: United Nations Statistics Division",
       x="Males",
       y="Females")+
theme_bw()
```

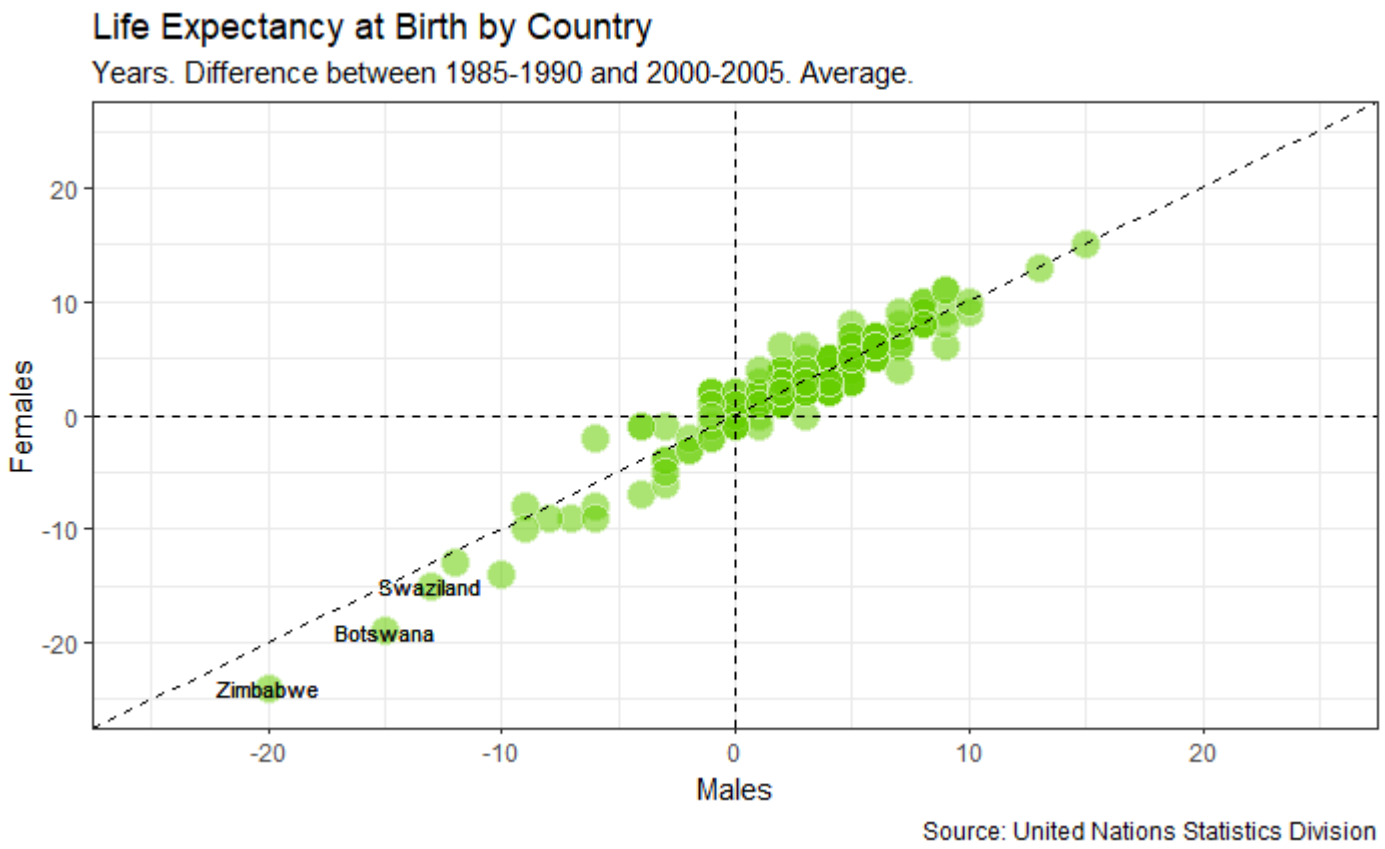


Highlighting remarkable countries II

As we did in the first plot, let's label some points. Concretely, we will point those three where the aggregated average life expectancy for men and women increased most and those three where decreased most in the period.

Hide

```
# Subseting data to obtain countries of interest
top <- subdata2 %>% arrange(diff_Male+diff_Female) %>% head(3)
bottom <- subdata2 %>% arrange(diff_Male+diff_Female) %>% head(3)
# Adding text to the previous plot to label countries of interest
ggplot(subdata2, aes(x=diff_Male, y=diff_Female, label=Country.or.Area), guide=FALSE)+
  geom_point(colour="white", fill="chartreuse3", shape=21, alpha=.55, size=5)+
  geom_abline(intercept = 0, slope = 1, linetype=2)+
  scale_x_continuous(limits=c(-25,25))+
  scale_y_continuous(limits=c(-25,25))+
  geom_hline(yintercept=0, linetype=2)+
  geom_vline(xintercept=0, linetype=2)+
  labs(title="Life Expectancy at Birth by Country",
       subtitle="Years. Difference between 1985-1990 and 2000-2005. Average.",
       caption="Source: United Nations Statistics Division",
       x="Males",
       y="Females")+
    geom_text(data=top, size=3)+
    geom_text(data=bottom, size=3)+
theme_bw()
```


Life Expectancy at Birth by Country
Years. Difference between 1985-1990 and 2000-2005. Average.
Females
Males
20
10
0
-10
-20
-20
-10
0
10
20
Swaziland
Botswana
Zimbabwe
Source: United Nations Statistics Division

Hide

```
NA
```