

STAT 261, Lab 3

HIV prevalence from WHO

- Estimated HIV prevalence was obtained from the `gapminder` website <https://www.gapminder.org/data/>
 - Estimated number of people living with HIV per 100 population of age group 15-49.
 - Original data source is the UNAIDS online database at <http://www.aidsinfoonline.org>
- A spreadsheet of the data, `HIVprev.csv`, is necessary for this lab.

We can read in these data as follows (we'll learn about reading in data later in STAT 260):

```
library(tidyverse) # you must have already installed the tidyverse package

## Warning: package 'tidyverse' was built under R version 4.0.2
## -- Attaching packages ----- tidyverse 1.3.0 --
## v ggplot2 3.3.2      v purrr   0.3.4
## v tibble  3.0.1      v dplyr  0.8.5
## v tidyr   1.1.2      v stringr 1.4.0
## v readr   1.3.1      v forcats 0.5.0
## Warning: package 'ggplot2' was built under R version 4.0.2
## Warning: package 'tidyr' was built under R version 4.0.2
## Warning: package 'readr' was built under R version 4.0.2
## Warning: package 'forcats' was built under R version 4.0.2
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
hiv <- read_csv("HIVprev.csv", stringsAsFactors = FALSE)
hiv <- select(hiv, Country, year, prevalence)
```

Take a look at the top and bottom few lines of raw data.

```
head(hiv)

##   Country year prevalence
## 1 Algeria 1990      0.06
## 2 Algeria 1991      0.06
## 3 Algeria 1992      0.06
## 4 Algeria 1993      0.06
## 5 Algeria 1994      0.06
## 6 Algeria 1995      0.06

tail(hiv)

##   Country year prevalence
## 1601 Zimbabwe 1995      25.1
```

```
## 1602 Zimbabwe 1996      26.2
## 1603 Zimbabwe 1997      26.5
## 1604 Zimbabwe 1998      26.3
## 1605 Zimbabwe 1999      25.7
## 1606 Zimbabwe 2000      24.8
```

```
summary(hiv)
```

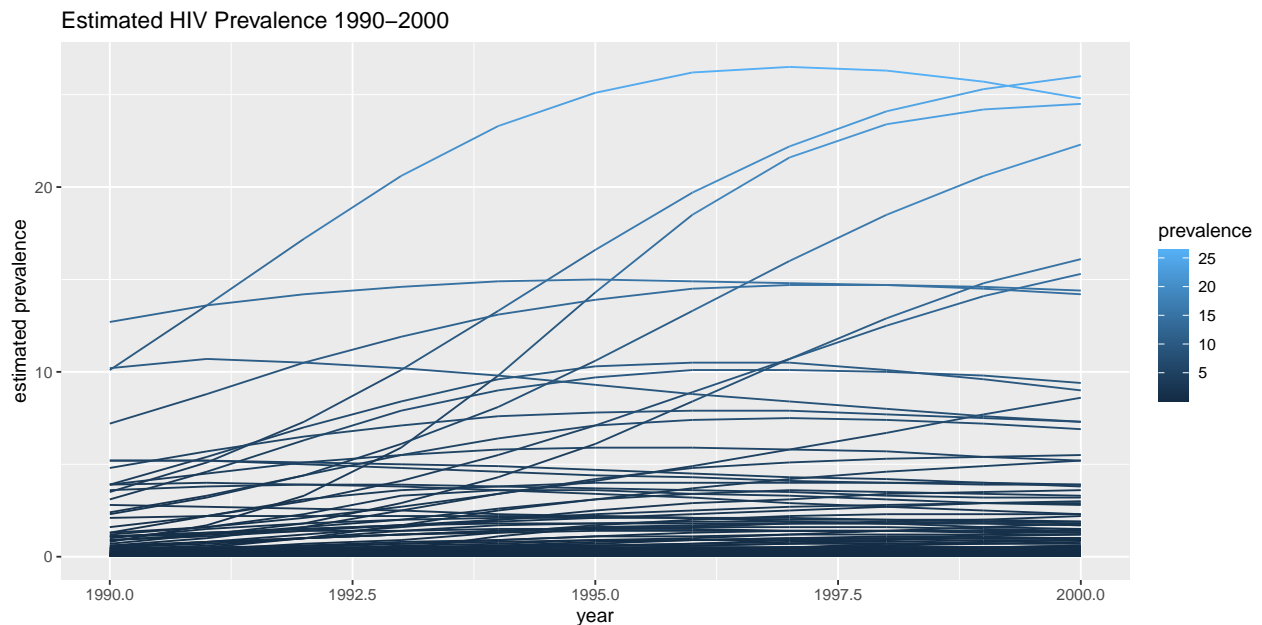
```
##      Country      year      prevalence
## Length:1606      Min.   :1990      Min.   : 0.060
## Class :character  1st Qu.:1992      1st Qu.: 0.060
## Mode  :character  Median :1995      Median : 0.200
##                               Mean  :1995      Mean  : 1.575
##                               3rd Qu.:1998      3rd Qu.: 1.100
##                               Max.   :2000      Max.   :26.500
```

Exercises:

1. Plot the time series of HIV prevalence by year for each country using `geom_line()`. Color the lines according to HIV prevalence. Add the title "Estimated HIV Prevalence 1990-2000" and change the y-axis label to "estimated prevalence".

Solution:

```
ggplot(hiv,aes(x=year,y=prevalence,group=Country,color=prevalence)) + geom_line() +
  labs(y="estimated prevalence",title="Estimated HIV Prevalence 1990-2000")
```

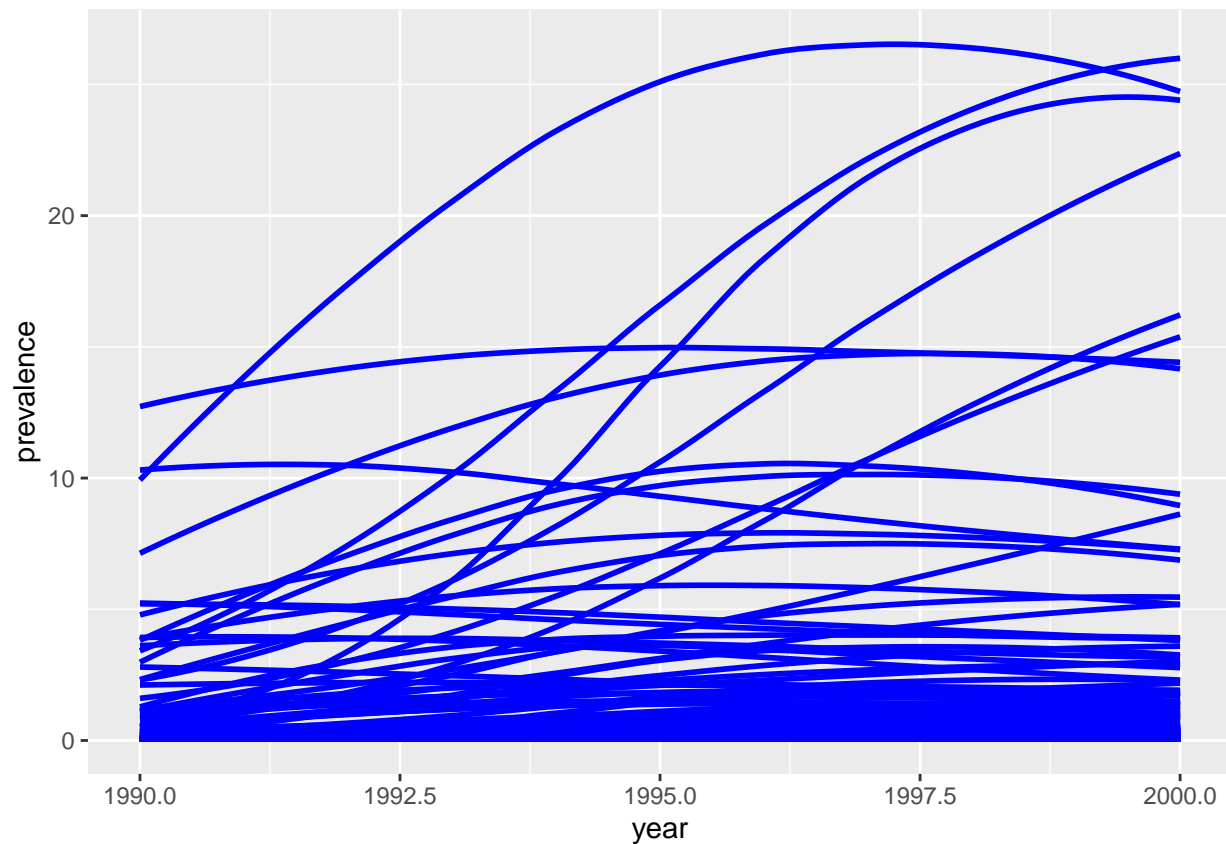


2. If you look closely at the previous plot you will notice that `geom_line()` draws “jagged” lines. This is because it draws a straight line between data points, as opposed to fitting a smooth curve. (To see this you can add a layer to the plot to include the points.) For this exercise, make a new time series plot. Instead of using `geom_line()`, fit and draw smoothers to represent the time series for each country. That is, plot *smooth* time series of HIV prevalence by year for each country (hint: use `geom_smooth()`). For this plot, make the drawn curves colored blue. (This plot should *not* include points, confidence bands, or any other superfluous details.)

Solution:

```
ggplot(hiv,aes(x=year,y=prevalence,group=Country)) + geom_smooth(colour="blue",se=FALSE)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



3. In the following code chunk we create a new dataset comprised of countries that had HIV prevalence greater than 10% in one or more of the years monitored (we will learn about this kind of “data wrangling” in future lectures of STAT 260).

```
cc <- c("Botswana", "Central African Republic", "Congo", "Kenya", "Lesotho", "Malawi",
        "Namibia", "South Africa", "Swaziland", "Uganda", "Zambia", "Zimbabwe")
hihiv <- filter(hiv, Country %in% cc)
```

Redo the time series plot from Exercise 1, with the following modifications. Color the time series for all but the countries in the `hihiv` data frame (i.e., those with high HIV prevalence) **grey** and with `alpha=0.3`. For the high-HIV-prevalence countries, color them **red**, also using `alpha=0.3`. Next, add two smoothers: (i) for all the data, i.e. all the countries in the `hiv` data frame, colored **black**, and (ii) for the countries with a high prevalence of HIV, i.e. those in the `hihiv` data frame, colored **red**.

Solution:

```
ggplot(hiv, aes(x=year, y=prevalence)) +
  geom_line(aes(group=Country), alpha=0.3, color="grey") +
  geom_line(data=hihiv, alpha=0.3, aes(group=Country), color="red") +
  geom_smooth(color="black") +
  geom_smooth(data=hihiv, color="red") +
  labs(y="estimated prevalence", title="Estimated HIV Prevalence 1990–2000")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

