R Exercise

Bob O'Hara

9 January 2019

Some Simple Exercises

We are going to look at the Olympic 100m times.

First, we need to read in the data. The data is on the

```
Olymp100m <- read.csv("https://www.math.ntnu.no/emner/ST2304/2019v/Week1/Olymp100m.csv")
```

First we can look at the data

```
str(Olymp100m)
```

```
## 'data.frame':    54 obs. of  3 variables:
##  $ Sex        : Factor w/ 2 levels "Men","Women": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Year       : int  1900 1904 1908 1912 1916 1920 1924 1928 1932 1936 ...
##  $ WinningTime: num  11 11 10.8 10.8 NA 10.8 10.6 10.8 10.3 10.3 ...
```

This says that this is a data frame (essentially a simple data table), with three variables: Sex, Year, and winning time are variables in the table. Sex is a factor, i.e. it has categories (called "levels"): obviously it has 2 levels, Men and Women. If we want to look at one variable, we can do this:

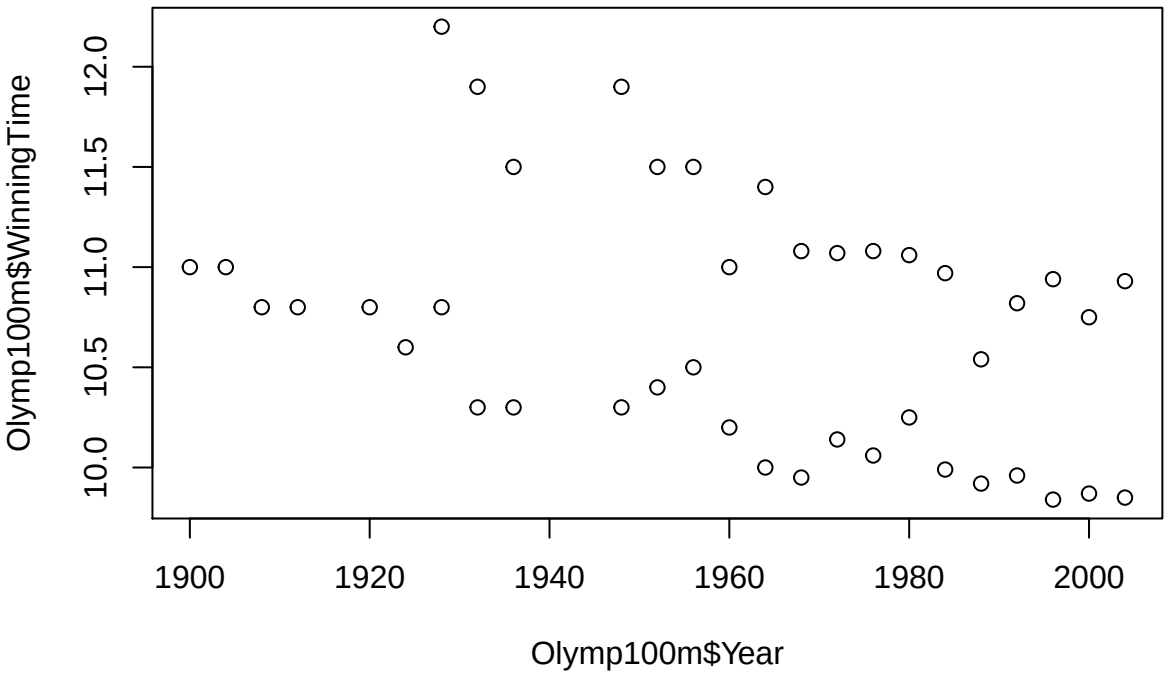
```
Olymp100m$WinningTime
```

```
##  [1] 11.00 11.00 10.80 10.80    NA 10.80 10.60 10.80 10.30 10.30    NA
## [12]    NA 10.30 10.40 10.50 10.20 10.00  9.95 10.14 10.06 10.25  9.99
## [23]  9.92  9.96  9.84  9.87  9.85    NA    NA    NA    NA    NA    NA
## [34]    NA 12.20 11.90 11.50    NA    NA 11.90 11.50 11.50 11.00 11.40
## [45] 11.08 11.07 11.08 11.06 10.97 10.54 10.82 10.94 10.75 10.93
```

Some of the values are given as "NA". This means that the value is missing (internally R codes it as a missing value, not with the letters "NA").

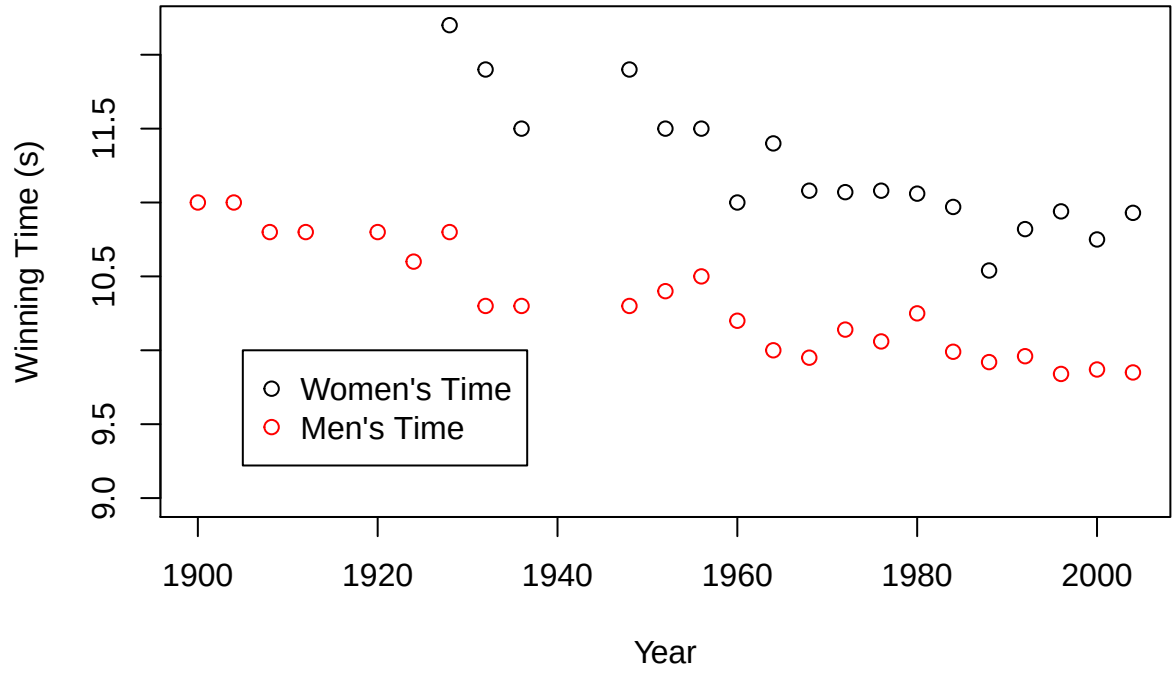
This is the simplest way to plot the data:

```
plot(Olymp100m$Year, Olymp100m$WinningTime)
```



The plot is OK to start with, but (for example) you cannot see which points are for men or women. So I improved the plot, and the code is below. I used a few options to improve the display (xlim=, ylim=, ylab=, xlab=, col=): try changing or removing them to see what the effect is. The second function, legend(), creates the legend. You have to run plot() first, otherwise R will complain because there is no plot to put a legend onto.

```
plot(Olymp100m$Year, Olymp100m$WinningTime, xlim=range(Olymp100m$Year),
     ylim=c(9, max(Olymp100m$WinningTime, na.rm = TRUE)),
     ylab="Winning Time (s)", xlab="Year", col=1+(Olymp100m$Sex=="Men"))
legend(1905, 10, c("Women's Time", "Men's Time"), col=c(1,2), pch=1)
```



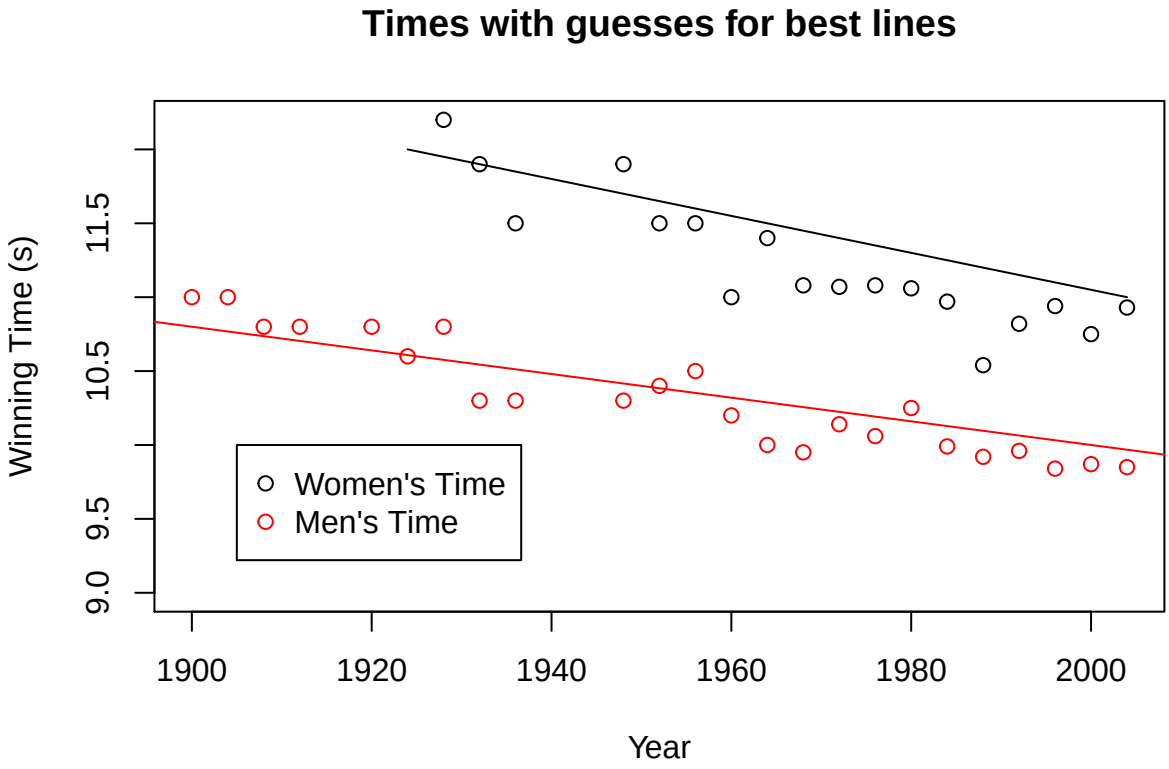
Some of the changes I made are simpler than others, for example the limits to the y-axis ('ylim') are written

as c(9, max(Olymp100m$WinningTime, na.rm = TRUE)). This becomes two numbers: 9 and 12.2. 9 is the lower limit, the upper limit is the maximum value (na.rm = TRUE tells R to remove any missing values before finding the maximum). R then draws the axis slightly outside these limits (if you want to use these limits exactly, you can tell R to do that too)

Finding a (visual) model for the data

You can, of course, print the graph out and draw lines on it, but you can also do this in R. There are a few ways to do this, here are a couple.

```
plot(Olymp100m$Year, Olymp100m$WinningTime, xlim=range(Olymp100m$Year),
     ylim=c(9, max(Olymp100m$WinningTime, na.rm = TRUE)),
     ylab="Winning Time (s)", xlab="Year", col=1+(Olymp100m$Sex=="Men"),
     main="Times with guesses for best lines")
lines(x=c(1924, 2004), y=c(12, 11)) # First way to add a straight line
abline(a=26, b=-0.008, col=2) # Second way to add a straight line
legend(1905, 10, c("Women's Time", "Men's Time"), col=c(1,2), pch=1)
```



The first, lines(), draws a line between the points: here I give two x coordinates (1924, 2004) and two y-coorinates (12, 11), so there are two points: the start and end of the line. The second, abline(), draws a line on the whole plot: here I give the intercept (26) and slope (-0.008).

Question for discussion: which of these ways of drawing the lines do you prefer? Why might someone prefer the other way? I can think of reasons for both ways, and I think the choice will depend on what you are trying to do.

Now try to find a better straight line by adjusting the lines and re-plotting them until you get ones that you think look right (we will be talking about better ways to do this later, of course). It will probably take you a few tries.

A couple of questions were raised about the data:

- whether the difference in times between men and women were decreasing
- whether the women's times had levelled out

Can you use what you have done to look at these questions? Or can you think of what to do to look at these questions in more detail? This is something we can come back to, and at the moment you should be a bit creative: even if you decide an idea doesn't work, deciding why it doesn't work can be really helpful.