Lab 1

Roxanne Li

2024-01-13

Read in some packages that we'll be using:

```
#install.packages("tidyverse")
library(tidyverse)
```

Read in mortality rates for Ontario:

```
dm <- read_table("https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt", skip = 2, col_types = "dcddd
head(dm)
```

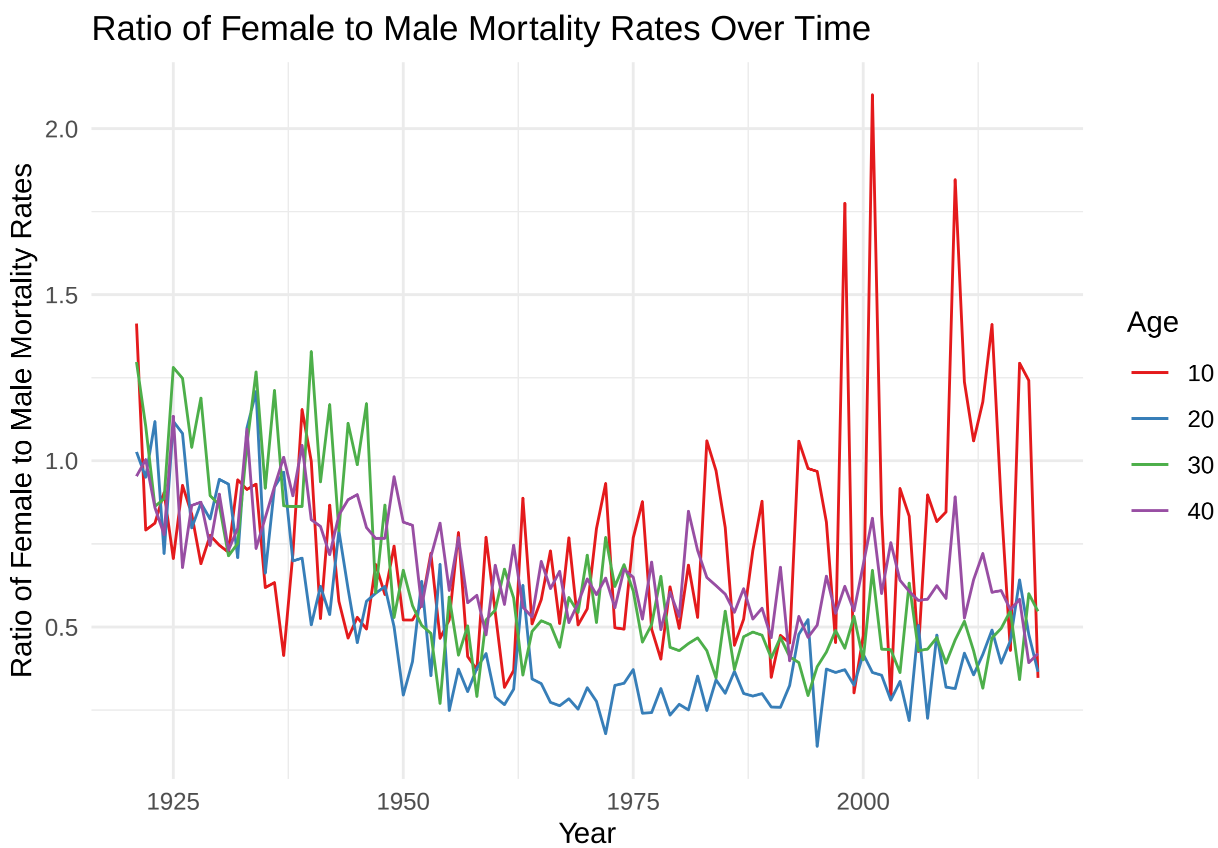
```
## # A tibble: 6 x 5
##    Year Age    Female    Male   Total
##   <dbl> <chr>   <dbl>   <dbl>   <dbl>
## 1  1921 0     0.0978  0.129   0.114
## 2  1921 1     0.0129  0.0144  0.0137
## 3  1921 2     0.00521 0.00737 0.00631
## 4  1921 3     0.00471 0.00457 0.00464
## 5  1921 4     0.00461 0.00433 0.00447
## 6  1921 5     0.00372 0.00361 0.00367
```

Q1

Plot the ratio of female to male mortality rates over time for ages 10,20,30 and 40 (different color for each age) and change the theme.

```
dm1 <- dm |>
  filter(Age==10|Age==20|Age==30|Age==40) |>
  mutate(mf_ratio = Female/Male)
```

```
dm1 |>
  ggplot(aes(x = Year, y = mf_ratio, color = Age)) +
  geom_line() +
  scale_color_brewer(palette = "Set1") +
  labs(title = "Ratio of Female to Male Mortality Rates Over Time",
       x = "Year",
       y = "Ratio of Female to Male Mortality Rates") +
  theme_minimal()
```



Q2

Find the age that has the lowest female mortality rate each year.

```
dm2 <- dm |>
  group_by(Year) |>
  slice(which.min(Female)) |>
  select(Year, Age, Female)
dm2
```

```
## # A tibble: 99 x 3
## # Groups:   Year [99]
##     Year Age     Female
##    <dbl> <chr>    <dbl>
##  1  1921 13     0.00176
##  2  1922 104    0
##  3  1923 105    0
##  4  1924 14     0.00140
##  5  1925 105    0
##  6  1926 11     0.000942
##  7  1927 9      0.00132
##  8  1928 9      0.00105
##  9  1929 10     0.00121
## 10  1930 13     0.00108
## # i 89 more rows
```

The table produced above shows the age that has the lowest female mortality rate each year.

Q3

Use the `summarize(across())` syntax to calculate the standard deviation of mortality rates by age for the Male, Female and Total populations.

```
dm |>
  group_by(Age) |>
  summarize(across(c(Male, Female, Total), sd, na.rm = TRUE))
```

```
## # A tibble: 111 x 4
##    Age       Male   Female    Total
##    <chr>    <dbl>    <dbl>    <dbl>
##  1 0     0.0330   0.0256   0.0294
##  2 1     0.00396  0.00352  0.00374
##  3 10    0.000561 0.000474 0.000509
##  4 100   0.138    0.0928   0.0729
##  5 101   0.158    0.125    0.0995
##  6 102   0.214    0.143    0.114
##  7 103   0.371    0.252    0.208
##  8 104   1.01     0.449    0.363
##  9 105   1.29     1.27     1.27
## 10 106   1.13     1.21     1.20
## # i 101 more rows
```

Q4

The Canadian HMD also provides population sizes over time (https://www.prdh.umontreal.ca/BDLC/data/ont/Population.txt). Use these to calculate the population weighted average mortality rate separately for males and females, for every year. Make a nice line plot showing the result (with meaningful labels/titles) and briefly comment on what you see (1 sentence). Hint: `left_join` will probably be useful here.

Reading the population file:

```
pop <- read_table("https://www.prdh.umontreal.ca/BDLC/data/ont/Population.txt",
                  skip = 2, col_types = "dcddd")
head(pop)
```

```
## # A tibble: 6 x 5
##    Year Age   Female   Male  Total
##   <dbl> <chr>  <dbl>  <dbl>  <dbl>
## 1  1921 0     30157. 31530. 61687.
## 2  1921 1     30391. 31319. 61711.
## 3  1921 2     30962. 31785. 62747.
## 4  1921 3     31306. 32031. 63336.
## 5  1921 4     31364. 32046. 63409.
## 6  1921 5     31175. 31847. 63021.
```

Left joining the mortality rate data with the population data and calculating population-weighted mortality rates for females and males:

```r
merged_data <- left_join(dm, pop, by=c("Year", "Age"), suffix=c(".dm", ".pop"), unmatched = "drop")
merged_data$Pop_Weighted_Female <- (merged_data$Female.dm + merged_data$Male.dm) *
  (merged_data$Female.pop / merged_data$Total.pop)
merged_data$Pop_Weighted_Male <- (merged_data$Female.dm + merged_data$Male.dm) *
  (merged_data$Male.pop / merged_data$Total.pop)
head(merged_data)
```

```
## # A tibble: 6 x 10
##    Year Age   Female.dm Male.dm Total.dm Female.pop Male.pop Total.pop
##   <dbl> <chr>     <dbl>   <dbl>    <dbl>      <dbl>    <dbl>     <dbl>
## 1  1921 0       0.0978  0.129    0.114       30157.   31530.    61687.
## 2  1921 1       0.0129  0.0144   0.0137      30391.   31319.    61711.
## 3  1921 2       0.00521 0.00737  0.00631     30962.   31785.    62747.
## 4  1921 3       0.00471 0.00457  0.00464     31306.   32031.    63336.
## 5  1921 4       0.00461 0.00433  0.00447     31364.   32046.    63409.
## 6  1921 5       0.00372 0.00361  0.00367     31175.   31847.    63021.
## # i 2 more variables: Pop_Weighted_Female <dbl>, Pop_Weighted_Male <dbl>
```

Calculating population-weighted average mortality rates for females and males, by year.

```r
pop_weighted_avg <- merged_data |>
  group_by(Year) |>
  summarize(across(c(Pop_Weighted_Female, Pop_Weighted_Male), mean, na.rm = TRUE))
head(pop_weighted_avg)
```

```
## # A tibble: 6 x 3
##    Year Pop_Weighted_Female Pop_Weighted_Male
##   <dbl>               <dbl>             <dbl>
## 1  1921              0.0928            0.170
## 2  1922              0.0821            0.103
## 3  1923              0.117             0.0988
## 4  1924              0.128             0.0726
## 5  1925              0.115             0.103
## 6  1926              0.188             0.0801
```

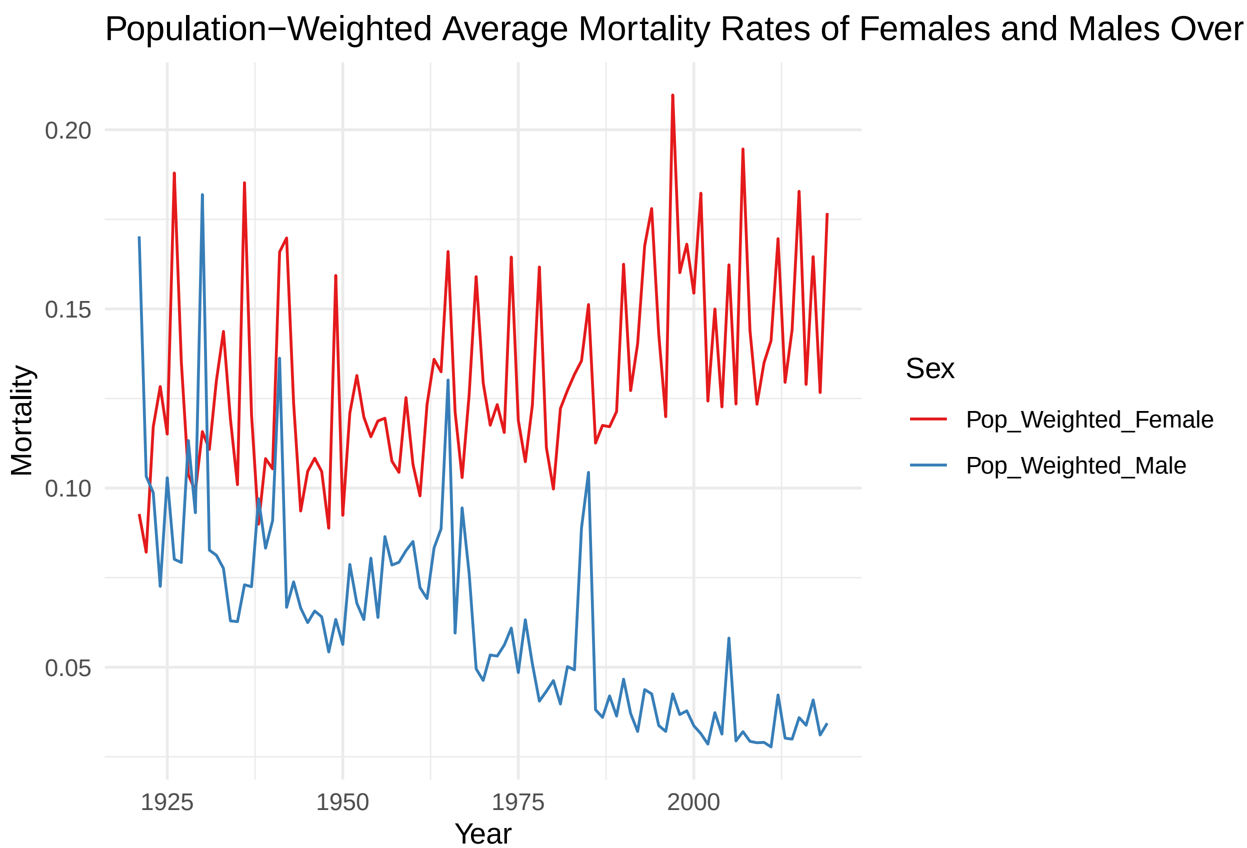
Turning the `pop_weighted_avg` into a pivot table for plotting:

```r
pop_weighted_avg_pivot <- pop_weighted_avg |>
  pivot_longer(Pop_Weighted_Female:Pop_Weighted_Male, names_to = "Sex",
               values_to = "Population_Weighted_Mortality")
head(pop_weighted_avg_pivot)
```

```
## # A tibble: 6 x 3
##    Year Sex                 Population_Weighted_Mortality
##   <dbl> <chr>                                       <dbl>
## 1  1921 Pop_Weighted_Female                        0.0928
## 2  1921 Pop_Weighted_Male                          0.170
## 3  1922 Pop_Weighted_Female                        0.0821
## 4  1922 Pop_Weighted_Male                          0.103
## 5  1923 Pop_Weighted_Female                        0.117
## 6  1923 Pop_Weighted_Male                          0.0988
```

Plotting the line plot for the population weighted average mortality rates of females and males over time:

```
pop_weighted_avg_pivot |>
  ggplot(aes(x = Year, y = Population_Weighted_Mortality, color = Sex)) +
  geom_line() +
  scale_color_brewer(palette = "Set1") +
  labs(title = "Population-Weighted Average Mortality Rates of Females and Males Over Time",
       x = "Year",
       y = "Mortality") +
  theme_minimal()
```



The population weighted average mortality rates show that over time, the rates for females have an upward trend and the rates for males have a downward trend.

Q5

Write down using appropriate notation, and run a simple linear regression with logged mortality rates as the outcome and age (as a continuous variable) as the covariate, using data for females aged less than 106 for the year 2000. Interpret the coefficient on age.

Getting the data:

```
data <- dm |>
  filter(Year==2000,  Age<106) |>
  select(c(Year, Age, Female))
head(data)
```

```
## # A tibble: 6 x 3
##    Year Age     Female
##   <dbl> <chr>    <dbl>
## 1  2000 0     0.00518
## 2  2000 1     0.000194
## 3  2000 10    0.000063
## 4  2000 100   0.413
## 5  2000 101   0.449
## 6  2000 102   0.442
```

We fit a linear regression model $log(y_i) = \beta_0 + \beta_1 x_i + \epsilon_i$ where y_i represents the female mortality rate, x_i represents the female age, and ϵ_i represents the random error.

```r
data$Age <- as.numeric(data$Age)
model <- lm(log(Female) ~ Age, data)
summary(model)
```

```
##
## Call:
## lm(formula = log(Female) ~ Age, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.37151 -0.01448  0.04905  0.07329  2.74426
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.007206   0.832088  -9.623 2.75e-05 ***
## Age          0.070634   0.009933   7.111 0.000192 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.391 on 7 degrees of freedom
## Multiple R-squared:  0.8784, Adjusted R-squared:  0.861
## F-statistic: 50.57 on 1 and 7 DF,  p-value: 0.0001918
```

The coefficient for `Age` is approximately 0.07. This means on average, if we keep everything else constant, then a female being one year older will be $e^{0.07} \approx 1.07$ times more likely to die in the year of 2000.