Homework 1

Yuhong Zhang

14/01/24

Table of contents


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

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.4.4     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

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

```
Warning: 494 parsing failures.
row    col               expected actual
108 Female no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1
109 Female no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1
110 Female no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1
110 Male   no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1
110 Total  no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1
```

```
... ...... .................... ...... .............................................
See problems(...) for more details.
```

```r
head(dm)
```

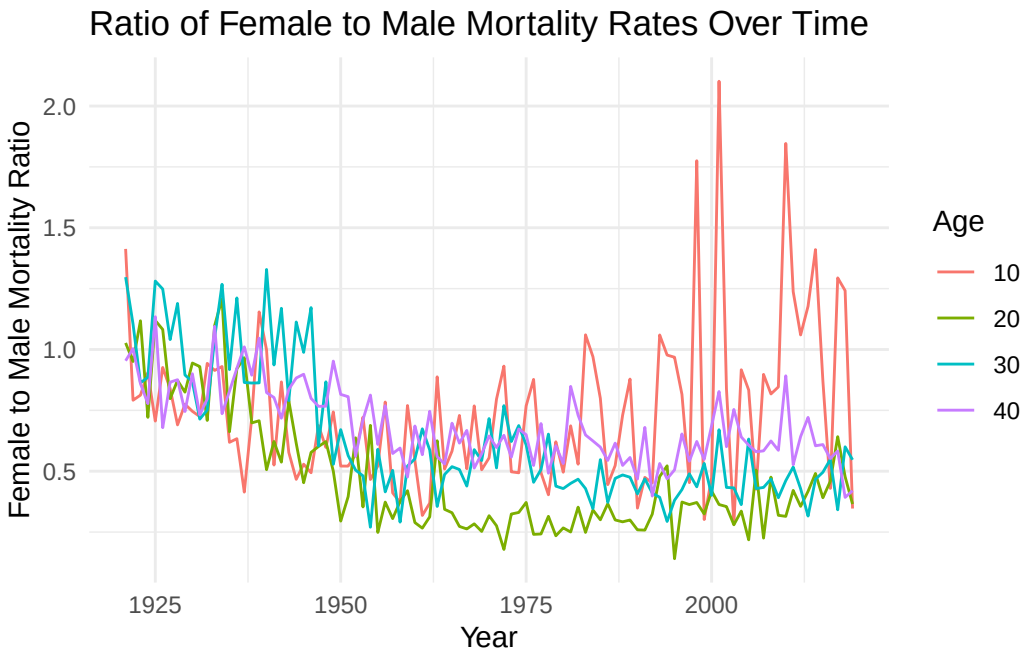
```
# A tibble: 6 x 5
   Year Age    Female    Male   Total
  <dbl> <chr>   <dbl>   <dbl>   <dbl>
1  1921 0     0.0978  0.129   0.114
2  1921 1     0.0129  0.0144  0.0137
3  1921 2     0.00521 0.00737 0.00631
4  1921 3     0.00471 0.00457 0.00464
5  1921 4     0.00461 0.00433 0.00447
6  1921 5     0.00372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

Lab Exercises

Make a new Quarto or R Markdown file to answer these questions, and push to your repository on Github (both the .qmd and pdf file) by Monday 9am. The file should be appropriately named, and in a folder in your repo called 'labs' or something similar.

1. 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

```r
dm|>
  filter(Age %in% c(10, 20, 30, 40)) |>
  select(Year:Male) |>
  mutate(Mortality_Ratio = Female / Male) |>
  pivot_longer(Female:Male, names_to = "Sex", values_to = "Mortality")|>
  ggplot(aes(x = Year, y = Mortality_Ratio, color = as.factor(Age))) +
  geom_line() +
  labs(title = "Ratio of Female to Male Mortality Rates Over Time",
       x = "Year",
       y = "Female to Male Mortality Ratio",
       color = "Age") +
  theme_minimal()
```



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

```
lowestfemalemortality <-dm  |>
  group_by(Year)|>
  filter(Female==min(Female,na.rm = TRUE))|>
  select(Year, Age, Female)
lowestfemalemortality
```

```
# A tibble: 171 x 3
# Groups:   Year [99]
    Year Age     Female
   <dbl> <chr>    <dbl>
 1  1921 13     0.00176
 2  1922 104    0
 3  1922 105    0
 4  1923 105    0
 5  1923 106    0
 6  1924 14     0.00140
 7  1925 105    0
 8  1925 106    0
 9  1926 11     0.000942
10  1927 9      0.00132
# i 161 more rows
```

Since there may be some age groups in a year that have the same female mortality rate and are the lowest,such as in 1922, age of 104 and 105 both obtain the lowest female mortality rate (0.00). Therefore, there are more than 99 elements in Year in this case.

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

```
dm$Age <- as.numeric(as.character(dm$Age))
```

```
Warning: NAs introduced by coercion
```

```
dm|>
  group_by(Age) |>
  summarize(across(2:4, ~ sd(., na.rm = TRUE)))
```

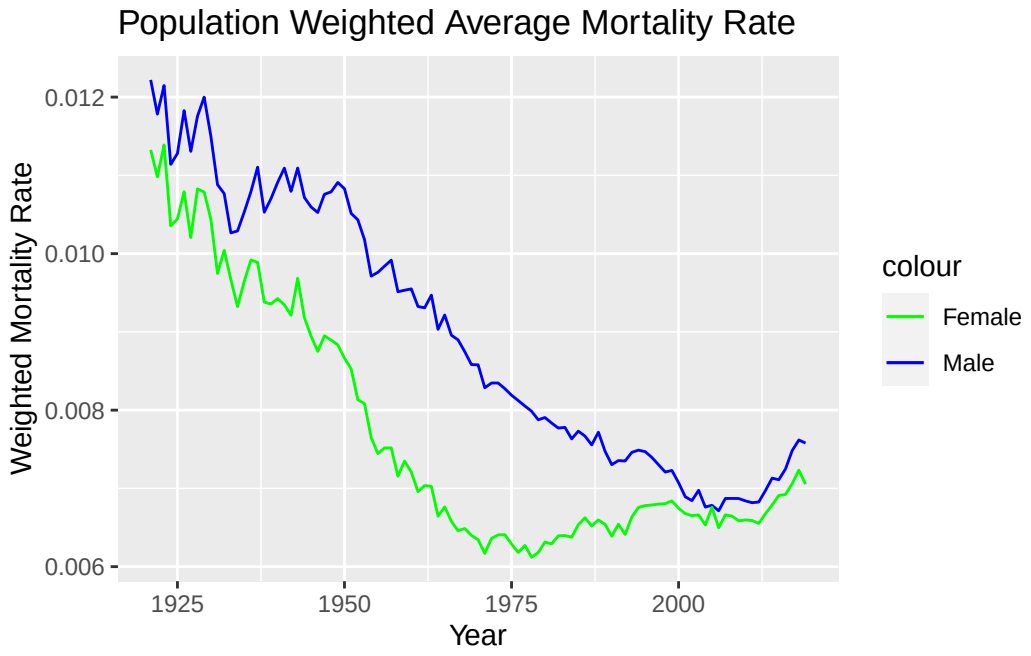
```
# A tibble: 111 x 4
     Age   Female     Male    Total
   <dbl>    <dbl>    <dbl>    <dbl>
 1     0 0.0256   0.0330   0.0294
 2     1 0.00352  0.00396  0.00374
 3     2 0.00154  0.00175  0.00164
 4     3 0.00113  0.00127  0.00120
 5     4 0.000925 0.000987 0.000947
 6     5 0.000748 0.000820 0.000776
 7     6 0.000631 0.000849 0.000731
 8     7 0.000590 0.000749 0.000664
 9     8 0.000496 0.000693 0.000590
10     9 0.000473 0.000604 0.000530
# i 101 more rows
```

4. The Canadian HMD also provides population sizes over time (https://www.prdh.umontreal.ca/BDLC/data 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.

```
da <- read_table("https://www.prdh.umontreal.ca/BDLC/data/ont/Population.txt",
                 skip = 2, col_types = "dcddd")
dm$Age <- as.character(dm$Age)
total<-
```

```
  left_join(da,dm, by = c("Year", "Age"))|>
  drop_na() |>
  group_by(Year) |>
  mutate(Weighted_Male_Mortality = Male.x * Male.y,
         Weighted_Female_Mortality = Female.x * Female.y) |>
  mutate(Avg_Male_Mortality = sum(Weighted_Male_Mortality,na.rm=TRUE)
         / sum(Male.x, na.rm=TRUE),
          Avg_Female_Mortality = sum(Weighted_Female_Mortality,na.rm=TRUE)
         / sum(Female.x,na.rm=TRUE))

total |>
  ggplot(aes(x = Year)) +
  geom_line(aes(y = Avg_Male_Mortality, color = "Male")) +
  geom_line(aes(y = Avg_Female_Mortality, color = "Female")) +
  labs(title = "Population Weighted Average Mortality Rate",
       x = "Year",
       y = "Weighted Mortality Rate") +
  scale_color_manual(values = c("Male" = "blue", "Female" = "green"))
```



From the plot, it is obvious that both male and female population weighted average mortality rates have generally decreased over the plot showed time period (mostly from 1921 to 2000) and the rate of woman is lower than male in general. After about 2008, there was slight increase in both female and male mortality rates, maybe because the technology development and

healthcare improvement, people can live longer, however, because the older population have higher mortality rates, development in health care could offset this effect, thus the mortality rates still much lower than before even if there is a small increase.

5. 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.

```r
dm$Age <- as.numeric(as.character(dm$Age))
sub<- dm |>
  filter(Year == 2000, Age < 106)
model <- lm(log(Female) ~ Age, data = sub)
summary(model)
```

```
Call:
lm(formula = log(Female) ~ Age, data = sub)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9692 -0.3194 -0.1341  0.2734  4.7993

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.062281   0.121345  -82.92   <2e-16 ***
Age           0.086891   0.001997   43.51   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6291 on 104 degrees of freedom
Multiple R-squared:  0.9479,    Adjusted R-squared:  0.9474
F-statistic:  1893 on 1 and 104 DF,  p-value: < 2.2e-16
```

From the result of linear regression, the coefficient of age means that when age increases by one unit (increase one year) and all other covariates are held constant (there is no other covariate in this case) , the mean of logged mortality rates will increase by 0.086891.