STA2201 HW1

Chenxi Liu 1010615050

2024-01-14

Lab Exercise 1

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     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 become errors
```

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

```
## Warning: 494 parsing failures.
## row    col               expected actual                                                  file
## 108 Female no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt'
## 109 Female no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt'
## 110 Female no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt'
## 110 Male   no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt'
## 110 Total  no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt'
## ... ...... ..................... ...... ..................................................
## See problems(...) for more details.
```

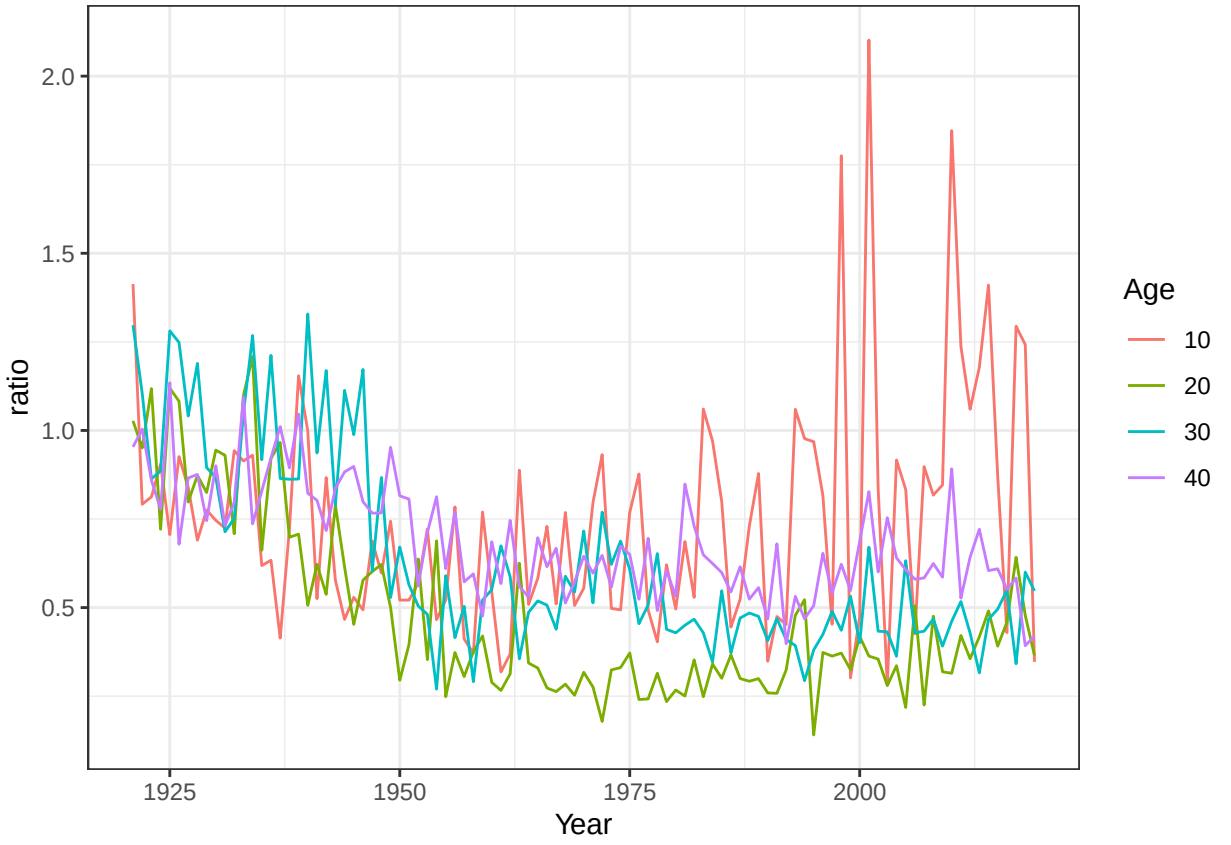
```
head(dm)
```

```
## # 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
```

Question1

```
plot1 <- dm %>%
  mutate(ratio = Female / Male) %>%
  filter(Age %in% seq(10, 40, by = 10)) %>%
  ggplot(aes(Year,ratio,color = Age)) +
  geom_line() +
  theme_bw()

plot1
```



Question2

```
result1 <- dm %>%
  group_by(Year) %>%
  filter(!is.na(Female) & !is.na(Male)) %>%
  arrange(Female) %>%
  slice(1) %>%
  ungroup()

result1
```

```
## # A tibble: 99 x 5
```

```
##     Year Age     Female    Male   Total
##    <dbl> <chr>    <dbl>   <dbl>   <dbl>
##  1  1921 13    0.00176  0.00184 0.00180
##  2  1922 104   0        1.20    0.794
##  3  1923 105   0        2.38    0.765
##  4  1924 14    0.00140  0.00186 0.00164
##  5  1925 105   0        4.18    1.05
##  6  1926 11    0.000942 0.00185 0.00140
##  7  1927 9     0.00132  0.00199 0.00166
##  8  1928 9     0.00105  0.00177 0.00142
##  9  1929 10    0.00121  0.00156 0.00139
## 10  1930 13    0.00108  0.00169 0.00139
## # i 89 more rows
```

Question3

```
library(dplyr)

result2 <- dm %>%
  group_by(Age) %>%
  summarize(across(c(Female, Male, Total), ~sd(., na.rm = TRUE)))

print(result2)
```

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

Question4

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

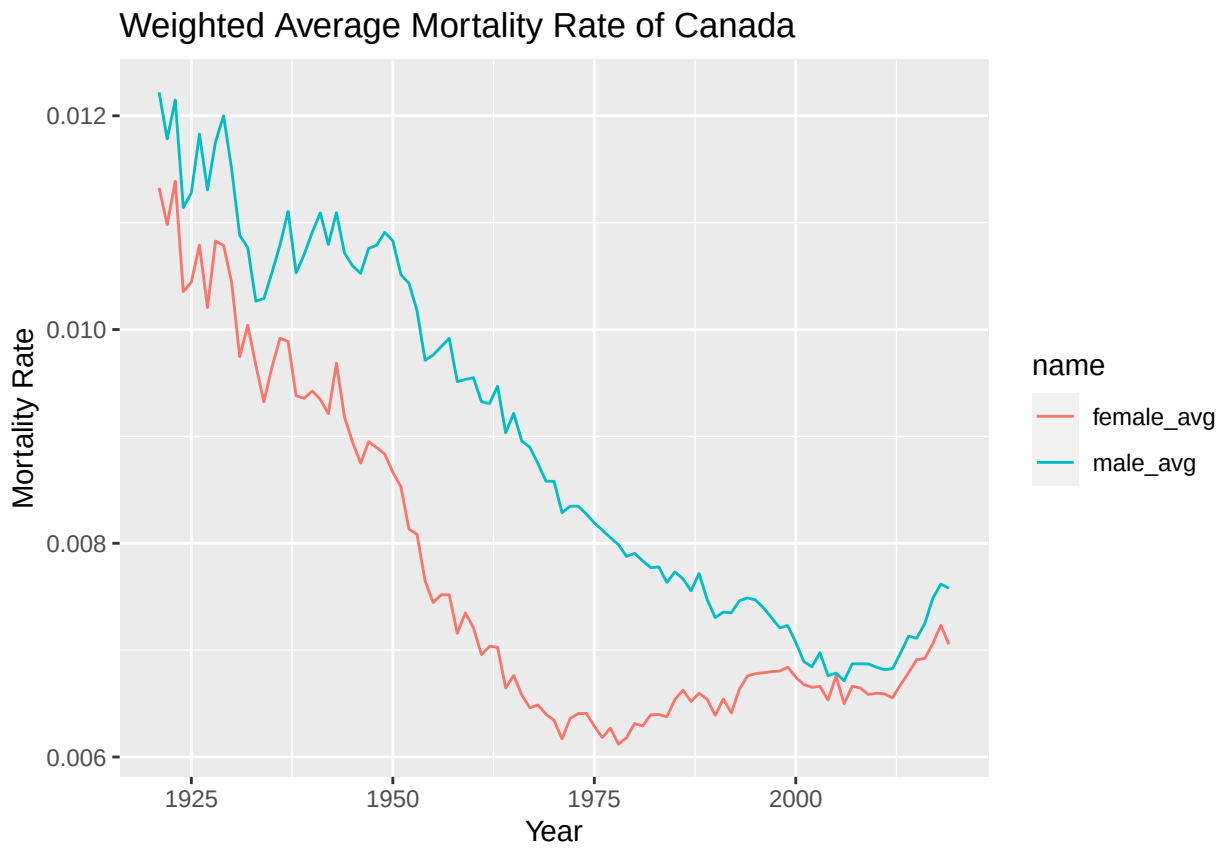
```
## # A tibble: 6 x 5
##    Year Age   Female   Male  Total
##   <dbl> <chr>  <dbl>  <dbl>  <dbl>
## 1  1921 0     30157. 31530. 61687.
## 2  1921 1     30391. 31319. 61711.
```

```
## 3  1921 2     30962. 31785. 62747.
## 4  1921 3     31306. 32031. 63336.
## 5  1921 4     31364. 32046. 63409.
## 6  1921 5     31175. 31847. 63021.
```

```
plot2 <- dm %>%
  select(-Total) %>%
  left_join(data %>%
              rename(pop_male = Male, pop_female = Female)) %>%
  drop_na() %>%
  group_by(Year) %>%
  summarise(female_avg = sum(Female*pop_female)/sum(pop_female),
            male_avg = sum(Male*pop_male)/sum(pop_male)) %>%
  pivot_longer(-Year) %>%
  ggplot(aes(Year, value, color = name)) +
  geom_line() +
  labs(y = 'Mortality Rate', x = 'Year', title = 'Weighted Average Mortality Rate of Canada')
```

```
## Joining with `by = join_by(Year, Age)`
```

```
plot2
```



As the plot shown above, we can see that the weighted average mortality rate of both female and male drop down dramatically till 1975, while the rate of female began to rise up from 1975 to 2000. What's more, the weighted average mortality rate of male is always higher than that of female throughout timeline.

Question5

```r
# Convert "Age" to numeric
dm$Age <- as.numeric(dm$Age)
```

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

```r
subset_data <- dm %>%
  filter(Year == 2000, Age < 106) %>%
  select(-c(Male,Total)) %>%
  drop_na()

str(subset_data)
```

```
## tibble [106 x 3] (S3: tbl_df/tbl/data.frame)
##  $ Year  : num [1:106] 2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
##  $ Age   : num [1:106] 0 1 2 3 4 5 6 7 8 9 ...
##  $ Female: num [1:106] 0.00518 0.000194 0.000187 0.000195 0.00008 0.000078 0.000078 0.00009 0.000076
```

```r
# Run a simple linear regression
model <- lm(log(Female) ~ Age, data = subset_data)

# Display the summary of the regression
summary(model)
```

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

According to the summary of model above, the estimate of coefficient of Age is 0.0869 with a significant p-value $< 2e * 10^6$, which is a positive number. It means that when the Age increases by 1 unit, the log of mortatility rate of female increases by 0.087 around, it indicates that mmortatility rate of female has a positive relationship with the grow of age.