Lab Exercises 1

Jongtaek Lee

2024-01-13

```
library(tidyverse)
```

```
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
```

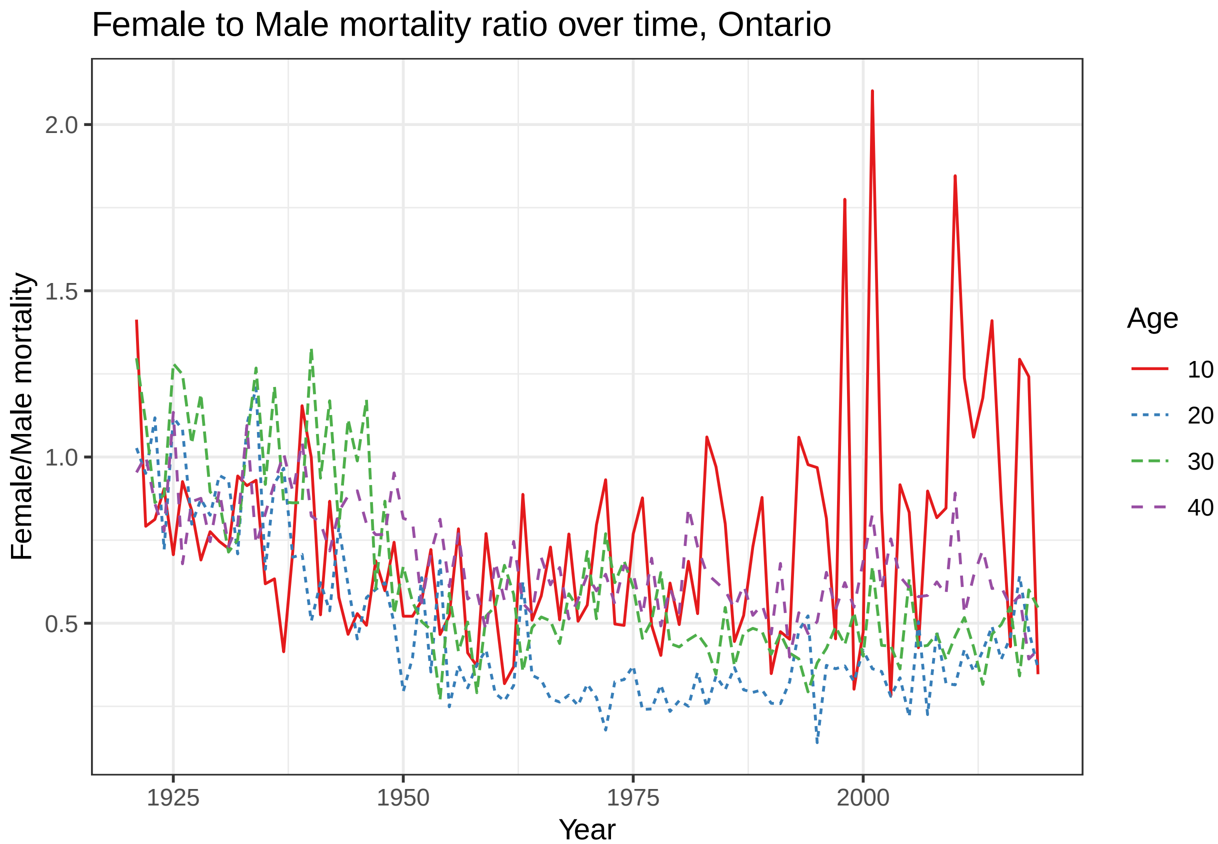
Question 1

```
dm_fm_ratio <- dm |>
                  mutate(fm_ratio = Female/Male) |>
                  filter(Age==10|Age==20|Age==30|Age==40) |>
                  select(Year:Age|fm_ratio)
dm_fm_ratio
```

```
## # A tibble: 396 x 3
##     Year Age   fm_ratio
##    <dbl> <chr>    <dbl>
##  1  1921 10       1.41
##  2  1921 20       1.03
##  3  1921 30       1.30
##  4  1921 40       0.954
##  5  1922 10       0.792
##  6  1922 20       0.951
##  7  1922 30       1.10
##  8  1922 40       1.00
##  9  1923 10       0.812
## 10  1923 20       1.12
## # i 386 more rows
```

```
dm_fm_ratio |>
  ggplot(aes(x=Year ,y=fm_ratio, color=Age, linetype=Age)) +
  geom_line() +
  scale_color_brewer(palette = "Set1") +
  theme_bw() +
```

```
  labs(title = "Female to Male mortality ratio over time, Ontario",
       y = "Female/Male mortality")
```



Question 2

```
dm |>
  select(Year:Age|Female) |>
  group_by(Year) |>
  summarize(Age[which.min(Female)])
```

```
## # A tibble: 99 x 2
##     Year `Age[which.min(Female)]`
##    <dbl> <chr>
##  1  1921 13
##  2  1922 104
##  3  1923 105
##  4  1924 14
##  5  1925 105
##  6  1926 11
##  7  1927 9
##  8  1928 9
##  9  1929 10
## 10  1930 13
## # i 89 more rows
```

Question 3

```
dm |>
  group_by(Age) |>
  summarize(across(Female:Total, sd, na.rm = TRUE)) |>
  arrange(as.numeric(Age))
```

```
## # 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 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
```

```
dm2 <- read_table("https://www.prdh.umontreal.ca/BDLC/data/ont/Population.txt", skip = 2, col_types = "c
head(dm2)
```

```
## # 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.
```

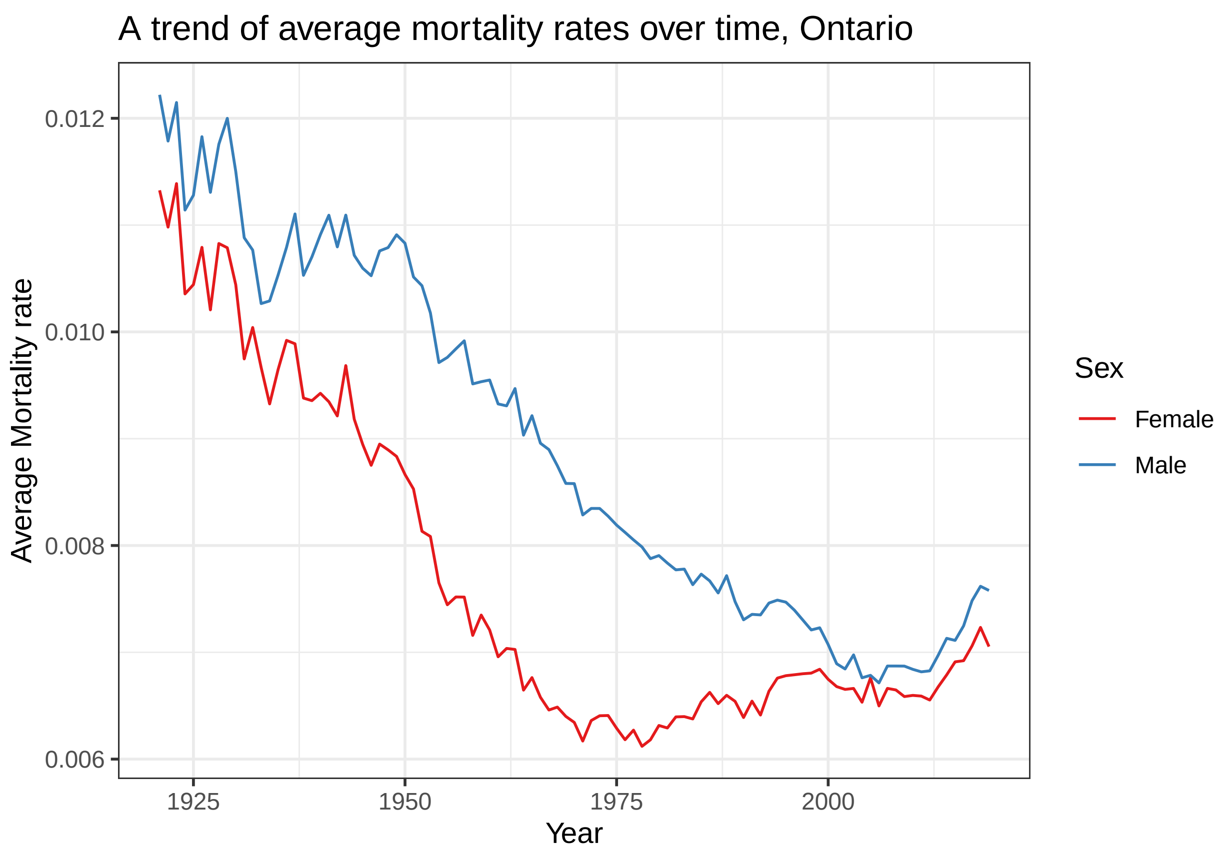
```
colnames(dm2) <- c("Year", "Age", "Female_pop", "Male_pop", "Total_pop")

dm_new <- dm |>
            left_join(dm2)
dm_new
```

```
## # A tibble: 10,989 x 8
##     Year Age    Female    Male   Total 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.
##  7  1921 6     0.00265 0.00393 0.00330     30808.   31466.    62274.
##  8  1921 7     0.00295 0.00351 0.00323     30295.   30922     61217.
##  9  1921 8     0.00237 0.00285 0.00262     29660.   30270.    59930.
## 10  1921 9     0.00198 0.00255 0.00227     28923    29494.    58417.
## # i 10,979 more rows
```

```r
dm_avg <- dm_new |>
  group_by(Year) |>
  summarize(Female = sum(Female*Female_pop, na.rm=TRUE)/sum(Female_pop, na.rm=TRUE),
            Male = sum(Male*Male_pop, na.rm=TRUE)/sum(Male_pop, na.rm=TRUE)) |>
  pivot_longer(Female:Male, names_to = "Sex", values_to = "Average_rate")
dm_avg
```

```
## # A tibble: 198 x 3
##     Year Sex    Average_rate
##    <dbl> <chr>         <dbl>
##  1  1921 Female       0.0113
##  2  1921 Male         0.0122
##  3  1922 Female       0.0110
##  4  1922 Male         0.0118
##  5  1923 Female       0.0114
##  6  1923 Male         0.0121
##  7  1924 Female       0.0104
##  8  1924 Male         0.0111
##  9  1925 Female       0.0104
## 10  1925 Male         0.0113
## # i 188 more rows
```

```r
dm_avg |>
  ggplot(aes(x=Year, y=Average_rate, color=Sex)) +
  geom_line() +
  scale_color_brewer(palette = "Set1") +
  labs(title = "A trend of average mortality rates over time, Ontario",
       y = "Average Mortality rate") +
  theme_bw()
```



Since 1975, the mortality rate for female started to keep increasing until 2000, however the one for male kept decreasing in the same period.

Question 5

```
y <- dm |>
        select(Year:Female) |>
        filter(Year == 2000, as.numeric(Age) < 106)
y
```

```
## # A tibble: 106 x 3
##      Year Age     Female
##     <dbl> <chr>    <dbl>
##   1  2000 0      0.00518
##   2  2000 1      0.000194
##   3  2000 2      0.000187
##   4  2000 3      0.000195
##   5  2000 4      0.00008
##   6  2000 5      0.000078
##   7  2000 6      0.000078
##   8  2000 7      0.00009
##   9  2000 8      0.000076
## 10  2000 9      0.000088
## # i 96 more rows
```

```r
y_data <- log(y$Female)
```

```r
model <- lm(y_data ~ as.numeric(y$Age), data=y)
summary(model)
```

```
##
## Call:
## lm(formula = y_data ~ as.numeric(y$Age), data = y)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9692 -0.3194 -0.1341  0.2734  4.7993
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -10.062281   0.121345  -82.92   <2e-16 ***
## as.numeric(y$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
```

Population regression model:

$log(Female_i) = \beta_{0i} + \beta_{1i} * Age_i + \epsilon_i$

Fitted regression model:

$log(Female_i) = -10.062281 + 0.086897 * Age_i$

The expected value of log of female mortality rate increases by 0.086891 for every unit increase of Age. Therefore, for a woman who gets one year older, her expected mortality rate will be exp(0.086891) = 1.090788 times of the current rate.