Lab1

Bingqing Li

```r
library(tidyverse)
library(ggplot2)
```

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

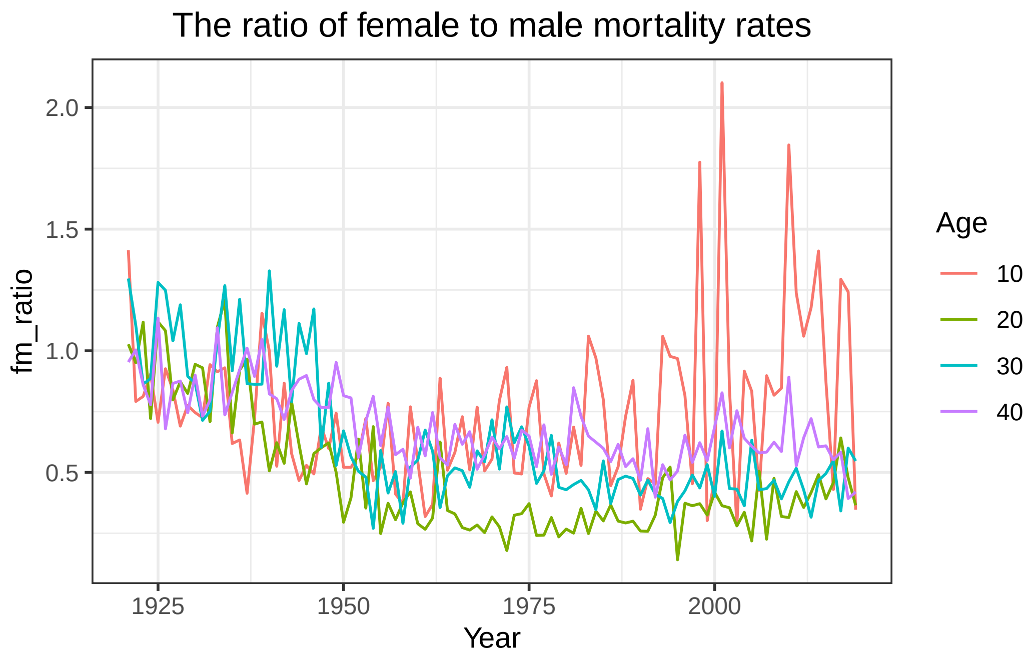
```
# A tibble: 6 x 5
   Year Age     Female    Male   Total
  <dbl> <chr>    <dbl>   <dbl>   <dbl>
1  1921 0      0.0978  0.129   0.114
2  1921 1      0.0129  0.0144  0.0137
3  1921 2      0.00521 0.00737 0.00631
4  1921 3      0.00471 0.00457 0.00464
5  1921 4      0.00461 0.00433 0.00447
6  1921 5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

Q1

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)) |>
  mutate(fm_ratio = Female/Male) |>
  ggplot(aes(x = Year, y = fm_ratio, color = Age)) +
  geom_line() +
  theme_bw() +
  labs(title = 'The ratio of female to male mortality rates') +
```

```
  theme(plot.title = element_text(hjust = 0.5))
```



Q2

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

```
dm |>
  group_by(Year) |>
  summarise(lf_age = Age[which.min(Female)])
```

```
# A tibble: 99 x 2
   Year lf_age
  <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
```

Q3

3. 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) |>
  summarise(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
```

Q4

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.

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

```
head(dp)
```

```
# 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.
```

```
dl <- left_join(dm, dp, by=c('Year', 'Age'))
colnames(dl) <- c('Year','Age', 'FemaleM', 'MaleM',
                  'TotalM', 'FemaleP', 'MaleP', 'TotalP' )
head(dl)
```

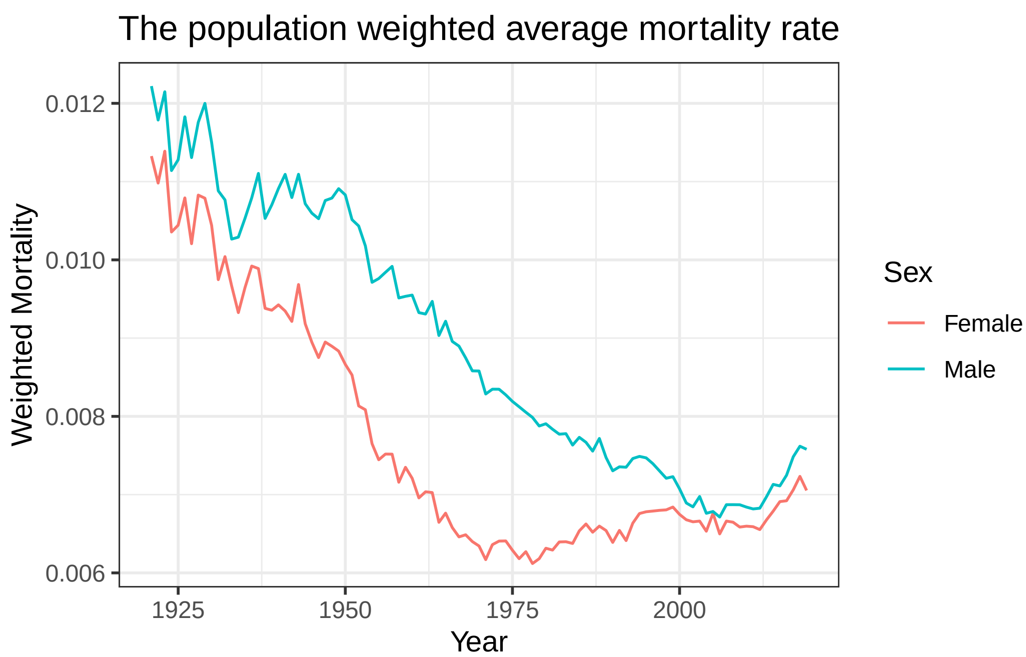
```
# A tibble: 6 x 8
   Year Age   FemaleM   MaleM  TotalM FemaleP  MaleP TotalP
  <dbl> <chr>   <dbl>   <dbl>   <dbl>   <dbl>  <dbl>  <dbl>
1  1921 0     0.0978  0.129   0.114    30157. 31530. 61687.
2  1921 1     0.0129  0.0144  0.0137   30391. 31319. 61711.
3  1921 2     0.00521 0.00737 0.00631  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.
```

```
dl1 <- dl |>
  group_by(Year) |>
  summarise(weighted_mean_F = sum(FemaleM*FemaleP,
                                  na.rm = TRUE) /sum(FemaleP, na.rm = TRUE),
            weighted_mean_M =  sum(MaleM*MaleP,
                                   na.rm = TRUE) /sum(MaleP, na.rm = TRUE) )|>
  pivot_longer(weighted_mean_F:weighted_mean_M, names_to = 'Sex',
               values_to = 'Weighted_Mean') |>
  mutate(Sex = case_when(
    Sex == 'weighted_mean_F' ~ 'Female',
    Sex == 'weighted_mean_M' ~ 'Male',
    TRUE ~ as.character(Sex)
  ))
```

```r
dl1
```

```
# A tibble: 198 x 3
    Year Sex    Weighted_Mean
   <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
dl1 |>
  ggplot(aes(x = Year, y = Weighted_Mean, color = Sex)) +
  geom_line() +
  theme_bw() +
  labs(y = 'Weighted Mortality',
       title = 'The population weighted average mortality rate') +
  theme(plot.title = element_text(hjust = 0.5))
```



The plot indicates that over the years, females have exhibited a lower average mortality rate compared to males, and the mortality rate for male and female shows a general downward trend.

Q5

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.

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

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

```
dm1 <- dm |>
  filter(Year == 2000 & Age < 106 & !is.na(Female)) |>
  select(Year, Age, Female)
dm1
```

```
# A tibble: 106 x 3
    Year   Age   Female
   <dbl> <dbl>    <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
```

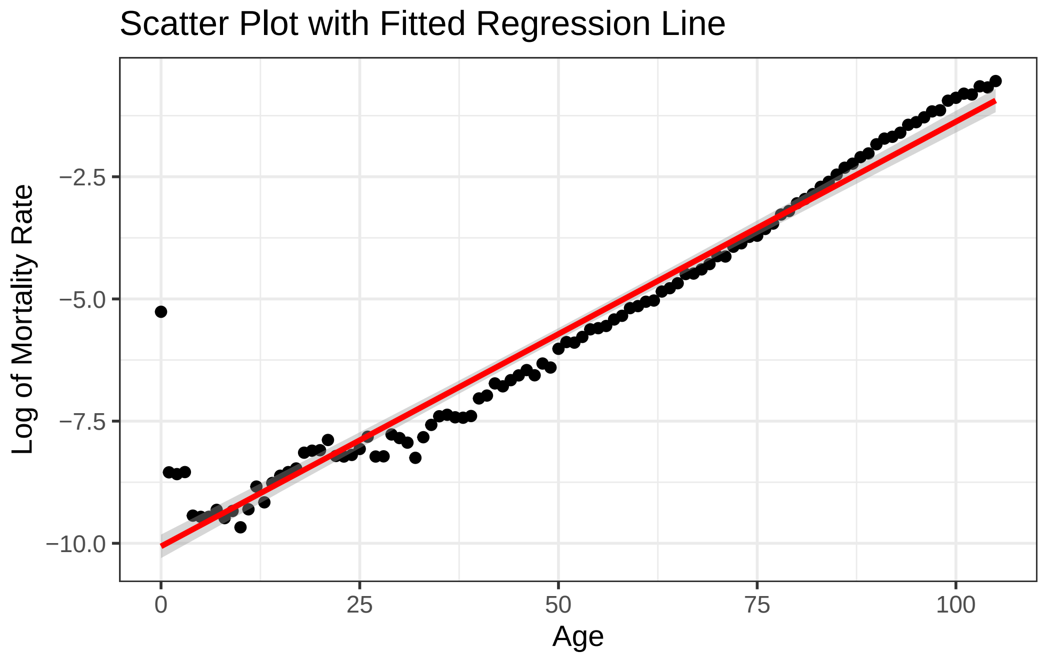
```r
model <-  lm(log(Female)~ Age, data = dm1)
coef(model)
```

```
 (Intercept)          Age
-10.06228123   0.08689126
```

```r
p <- ggplot(dm1, aes(x = Age, y = log(Female))) +
  geom_point() +
  geom_smooth(method = "lm", color = "red") +
  labs(
    x = "Age",
    y = "Log of Mortality Rate",
    title = "Scatter Plot with Fitted Regression Line"
  ) +
  theme(plot.title = element_text(hjust = 0.5))+
  theme_bw()

p
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Because

$$log(Female's\ Mortality_i) = 0.087 \times Age_i - 10.062,$$

then

$$Female's\ Mortality_i = e^{0.087 \times Age_i - 10.062}.$$

Thus, for females aged less than 106 for the year 2000, the mortality rate increases with age. Expected value change $e^{0.087}$ in Mortality rate with one unit increase in Age.