Applied_Stat_2_Lab_1

Rudraharsh Tewary

2024-01-15

Lab 1

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.2
```

```
## Warning: package 'readr' was built under R version 4.3.2
```

```
## Warning: package 'forcats' was built under R version 4.3.2
```

```
## -- 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   3.4.4     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 become errors
```

```r
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.
```

```r
dm
```

```
## # A tibble: 10,989 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
## 7  1921 6     0.00265 0.00393 0.00330
## 8  1921 7     0.00295 0.00351 0.00323
## 9  1921 8     0.00237 0.00285 0.00262
## 10 1921 9     0.00198 0.00255 0.00227
## # i 10,979 more rows
```

1)

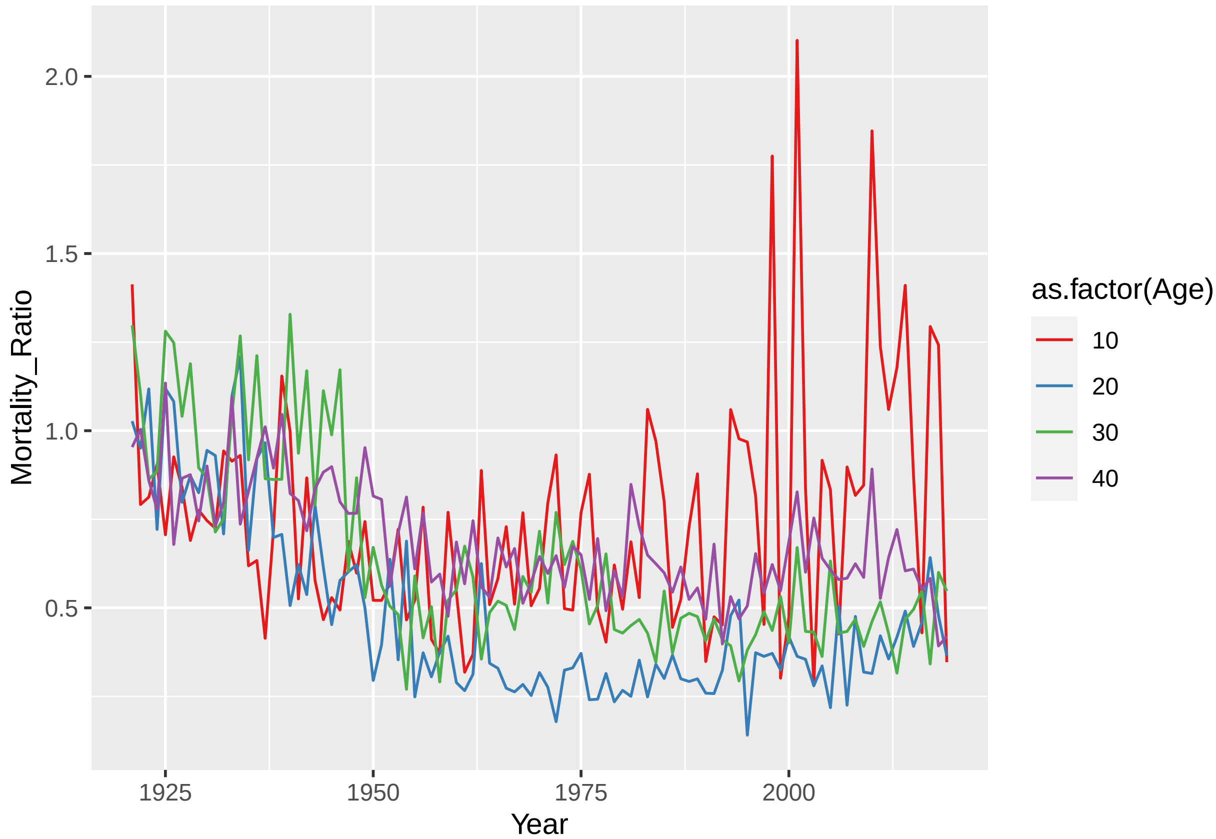
So, we need to plot the male to female mortality rates for ages of 10,20,30 and 40, we do it through the following code Block :-

```
d_to_plot <- 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")
d_to_plot
```

```
## # A tibble: 792 x 5
##     Year Age   Mortality_Ratio Sex    Mortality
##    <dbl> <chr>           <dbl> <chr>      <dbl>
##  1  1921 10              1.41  Female   0.00239
##  2  1921 10              1.41  Male     0.00169
##  3  1921 20              1.03  Female   0.00298
##  4  1921 20              1.03  Male     0.00290
##  5  1921 30              1.30  Female   0.00486
##  6  1921 30              1.30  Male     0.00375
##  7  1921 40              0.954 Female   0.00618
##  8  1921 40              0.954 Male     0.00648
##  9  1922 10              0.792 Female   0.00159
## 10  1922 10              0.792 Male     0.00201
## # i 782 more rows
```

Now, we get the following plot

```
d_to_plot |>
  ggplot(aes(x = Year, y = Mortality_Ratio, color = as.factor(Age))) +
  geom_line() +
  scale_color_brewer(palette = "Set1")
```



2)

We can find the age with the lowest mortality rate through the following code :-

```
lowest_mortality_age <- dm|>
  group_by(Year)|>
  arrange(Female)|>
  slice(1) |>
  select(Year, Age, Female)
lowest_mortality_age
```

```
## # 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
```

3)

We can find the standard deviation through the following code :-

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

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

```r
std_dev <- dm |>
  group_by(Age) |>
  summarise(across(2:4,sd,na.rm=TRUE))
```

```
## Warning: There was 1 warning in `summarise()`.
## i In argument: `across(2:4, sd, na.rm = TRUE)`.
## i In group 1: `Age = 0`.
## Caused by warning:
## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.
## Supply arguments directly to `.fns` through an anonymous function instead.
##
##   # Previously
##   across(a:b, mean, na.rm = TRUE)
##
##   # Now
##   across(a:b, \(x) mean(x, na.rm = TRUE))
```

```r
std_dev
```

```
## # 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
```

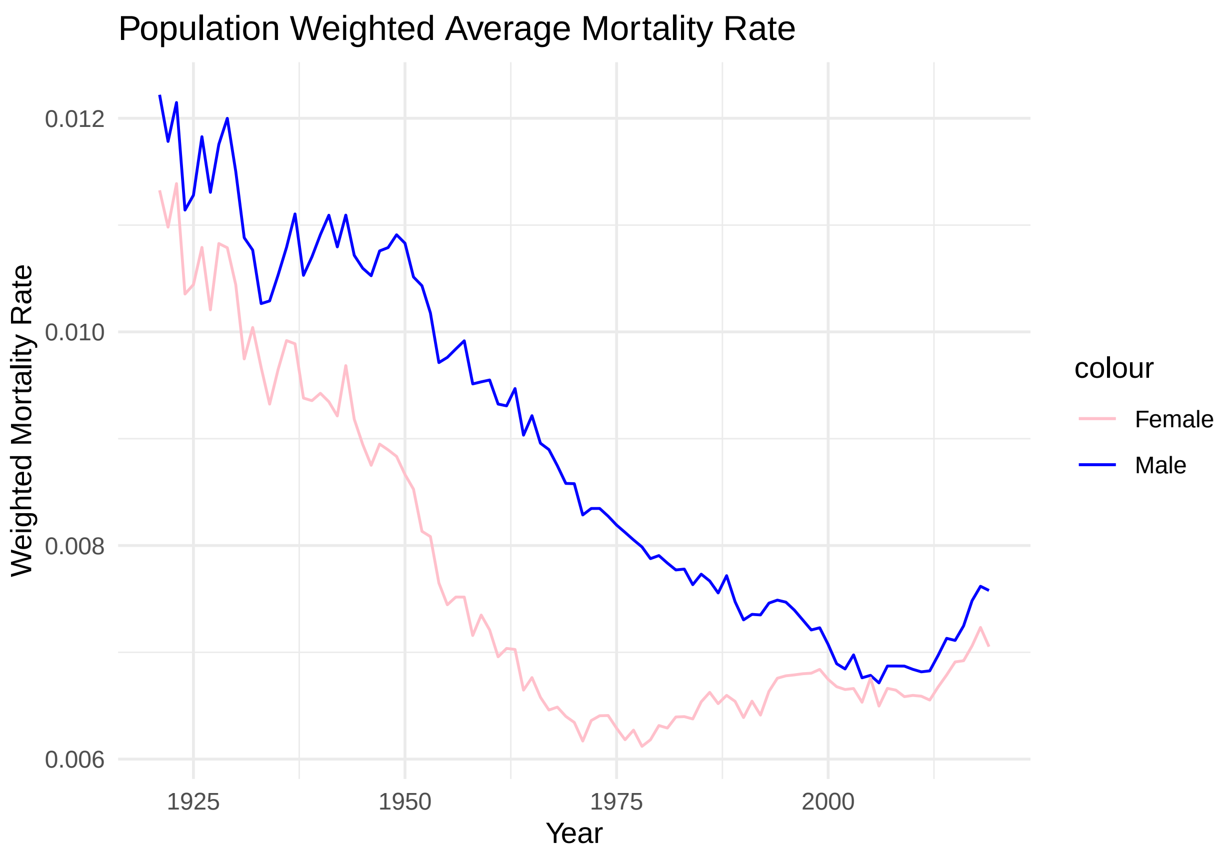
4)

We get our new table as :-

```r
dl <- read_table("https://www.prdh.umontreal.ca/BDLC/data/ont/Population.txt", skip = 1, col_types = "d
dm$Age<- as.character(dm$Age)
total<-left_join(dm,dl, by = c("Year","Age"))|>
  group_by(Year) |>
  drop_na() |>
  summarize(Avg_Male_Mortality = weighted.mean(Male.x, w=Male.y, na.rm = TRUE),
            Avg_Female_Mortality = weighted.mean(Female.x, w=Female.y, na.rm = TRUE))
total
```

```
## # A tibble: 99 x 3
##     Year Avg_Male_Mortality Avg_Female_Mortality
##    <dbl>              <dbl>                <dbl>
##  1  1921             0.0122               0.0113
##  2  1922             0.0118               0.0110
##  3  1923             0.0121               0.0114
##  4  1924             0.0111               0.0104
##  5  1925             0.0113               0.0104
##  6  1926             0.0118               0.0108
##  7  1927             0.0113               0.0102
##  8  1928             0.0118               0.0108
##  9  1929             0.0120               0.0108
## 10  1930             0.0115               0.0104
## # i 89 more rows
```

```r
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" = "pink")) +
  theme_minimal()
```



When we look at the plot, we see that the Weighted Male mortality rate was higher than the female mortality rate between the years of 1925 to 2000 because of multiple reasons, some deaths being caused due to males doing more risky/unsafe jobs which could have worker casualties and, a big contributor to male deaths would also be the world wars, where 66,000 canadians lost their lives in World War 1 and over 45,000 canadians died in World War 2. ### 5)

We will run the linear regression using the following code snippet

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

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

```
## # i 96 more rows
```

```r
model <- lm(log(Female) ~ Age, data = lm_table)
summary(model)
```

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

Here, we have a regression coefficient of 0.086891 for age. Noting the fact that the female mortality rate in our model is logged, this implies that keeping everything else constant, for every 1 unit increase in Age of a female, we would see an 8.6891% increase in the mortality rate.