Week 3 Lab

SOC6708 ADA

Alysia De Melo

```
library(tidyverse)
library(here)
library(readxl)
library(janitor)
```

```
d_male <- suppressWarnings(read_xlsx(here
                                     ("data/WPP2024_POP_F01_2_POPULATION_SINGLE_AGE_MALE.xlsx
d_male$sex <- "Male"
d_male <- d_male |> drop_na(Year)
d_female <- suppressWarnings(read_xlsx(here("data/WPP2024_POP_F01_3_POPULATION_SINGLE_AGE_FE
d_female$sex <- "Female"
d_female <- d_female |> drop_na(Year)

d <- rbind(d_male, d_female)
rm(d_male, d_female)

d <- d  |>
  clean_names() |>
  select(region_subregion_country_or_area, iso3_alpha_code, year, x0:sex) |>
  rename(region = region_subregion_country_or_area) |>
  mutate(across(x0:x100, as.numeric))

d_male <- suppressWarnings(read_xlsx(here("data/WPP2024_MORT_F01_2_DEATHS_SINGLE_AGE_MALE.xls
d_male$sex <- "Male"
d_male <- d_male |> drop_na(Year)
d_female <- suppressWarnings(read_xlsx(here("data/WPP2024_MORT_F01_3_DEATHS_SINGLE_AGE_FEMALE
d_female$sex <- "Female"
d_female <- d_female |> drop_na(Year)

dm <- rbind(d_male, d_female)
rm(d_male, d_female)
```

```r
dm <- dm  |>
  clean_names() |>
  select(region_subregion_country_or_area, iso3_alpha_code, year, x0:sex) |>
  rename(region = region_subregion_country_or_area) |>
  mutate(across(x0:x100, as.numeric))

d_long <- d  |>
  pivot_longer(x0:x100, names_to = "age", values_to = "pop") |>
  mutate(age = as.numeric(str_remove(age, "x")))

dm_long <- dm  |>
  pivot_longer(x0:x100, names_to = "age", values_to = "deaths") |>
  mutate(age = as.numeric(str_remove(age, "x")))

# join these two tibbles and calculate rates

asmr <- d_long |>
  left_join(dm_long) |>
  mutate(mx = deaths/pop)
```

Exercise

Decompose the difference in CDRs between USA and Japan in the year 2023. Is the majority of the difference due to age structure or mortality?

The decomposition of CDRs between the United States and Japan in 2023 shows a total difference of -0.00359. Japan's higher CDR is mainly attributable to its older age structure, as the age total is larger in absolute value compared to the mortality rate

```r
asmr |>
  filter(region == "United States of America",year==2023)  |>
  select(sex, age, pop, mx) |>
  rename(pop_usa = pop, mx_usa = mx) |>
  left_join(asmr |>
              filter(region == "Japan",year==2023)  |>
              select(sex, age, pop, mx) |>
              rename(pop_japan = pop, mx_japan = mx) ) |>
  mutate(prop_usa = pop_usa/sum(pop_usa),
         prop_japan = pop_japan/sum(pop_japan)) |>
  mutate(rate_diff = mx_usa - mx_japan,
         prop_diff = prop_usa - prop_japan) |>
```

```
  mutate(ave_rate = (mx_usa+mx_japan)/2,
         ave_prop = (prop_usa+prop_japan)/2) |>
  mutate(age_contr = prop_diff*ave_rate,
         rate_contr = rate_diff*ave_prop) |>
  summarize(age_total_contr = sum(age_contr),
            rate_total_contr = sum(rate_contr)) |>
  mutate(total_diff = age_total_contr+rate_total_contr)
```

```
# A tibble: 1 x 3
  age_total_contr rate_total_contr total_diff
            <dbl>            <dbl>      <dbl>
1        -0.00736          0.00377   -0.00359
```

Exercise

Now fit a Gompertz model to male mortality rates form age 40 in every year. Plot the estimated alpha and beta coefficients in a scatter plot, color the points by year. Comment on what you observe.

The graph shows that from 1921 until about 2007, individuals appear to be living longer, with death occurring later in life over time. During this period, improvements in mortality are steady and follow a clear pattern. From 2007 to 2023, this pattern appears to slow, as alpha no longer continues its earlier trend and instead levels off. This suggests that in more recent years, compared to the early 2000s, delaying death to later ages has stalled.

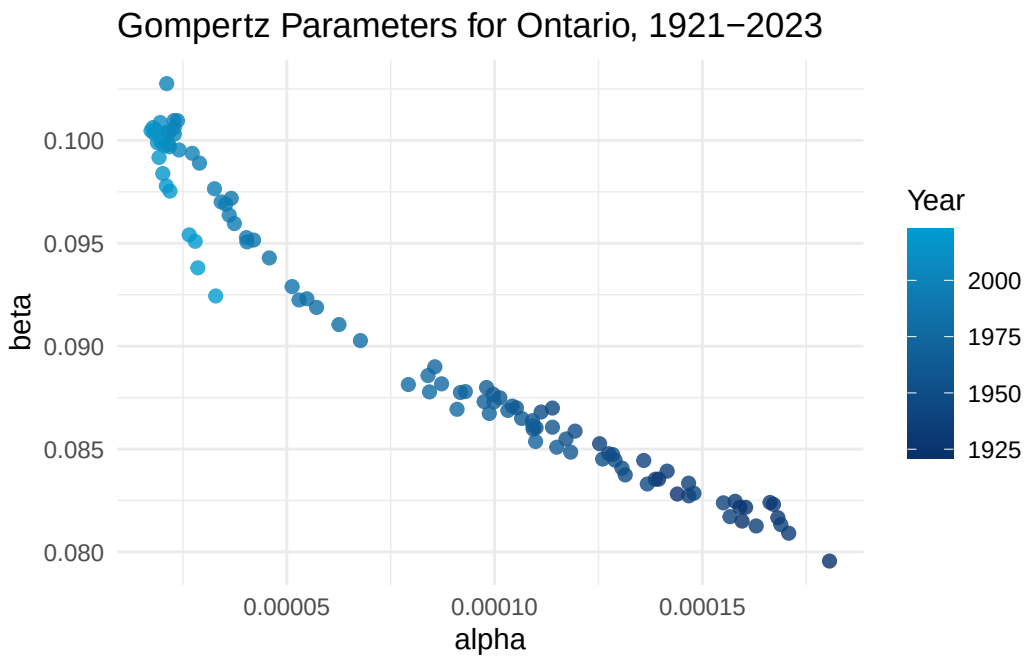
```
dm <- read_table("https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt", skip = 2, col_type
head(dm)
```

```
# 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
```

```r
#reproduce slide 20
dm <- dm |>
  clean_names() |>
  mutate(age = as.numeric(age))
```

```r
dm_to_fit <- dm |>
  filter(age>39, age<99) |>
  select(year, age, male)

coefs <- tibble()

years <- unique(dm_to_fit$year)

for(i in 1:length(years)){
  this_df <- dm_to_fit |>
    filter(year == years[i])

  this_mod <- lm(log(male)~age, data = this_df)

  these_coefs <-
    tibble(
      year  = years[i],
      alpha = exp(coef(this_mod)[1]),
           beta = coef(this_mod)[2])

  coefs <- bind_rows(coefs, these_coefs)

}
```

```r
ggplot(coefs, aes(x = alpha, y = beta, color = year)) +
  geom_point(size = 2, alpha = 0.8) +
  scale_color_gradient(low = "#08306b", high = "#009DD1") +
  labs(
    title = "Gompertz Parameters for Ontario, 1921-2023",   color = "Year"
  ) +
  theme_minimal()
```

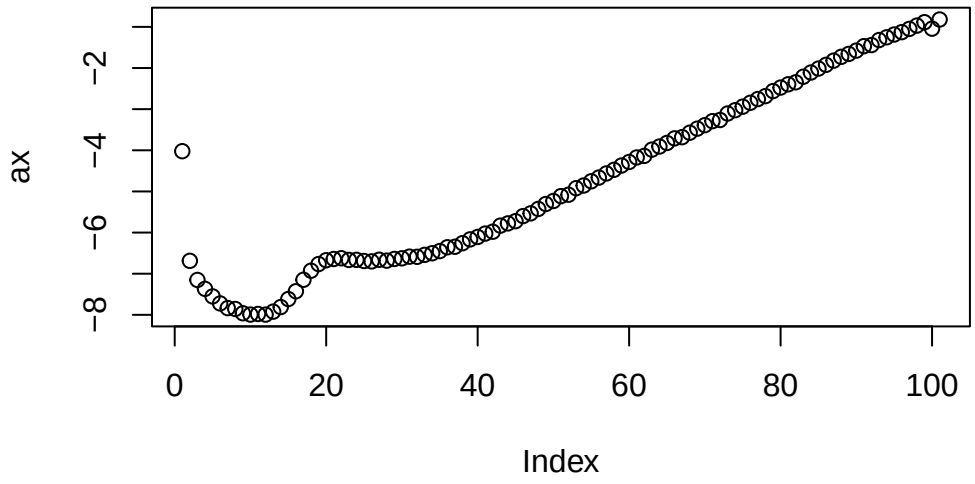


Exercise

Repeat the lee-carter model fitting exercise but just use mortality rates from 1970. Does this change the estimated rates? Does it do a better or worse job, or does it depend on the year?

Repeating the Lee–Carter fitting exercise using only mortality rates from 1970 onward does change the estimated parameters and fitted rates. This happens because the age-specific parameters of alpha and beta are now estimated from a shorter and more recent time period. As a result, the model is less influenced by earlier mortality patterns and these may no longer be relevant and so this may improve the models fit. However, the improvement is mixed and depends on the year being examined. The previous Lee-Carter only focused on three years post 1970. For 1981, the original model using the full 1921–2021 data set already fits very well. The fitted curve closely matches the observed log mortality rates, and the post-1970 data fits the data similarly. The 2011 graph performs similarly well in both models however both models have some discrepancies for those aged around 3 to around age 13. In 2021, both models again have some differences between points and the fitted line around ages 25–50, however here the full-sample model shows larger deviations from the observed rates compared to the post-1970 model. Overall, using only post-1970 data does not uniformly improve the fit. The benefits are modest and do not appear in every year. Thus, whether the restricted model performs better or worse depends on the year being evaluated.

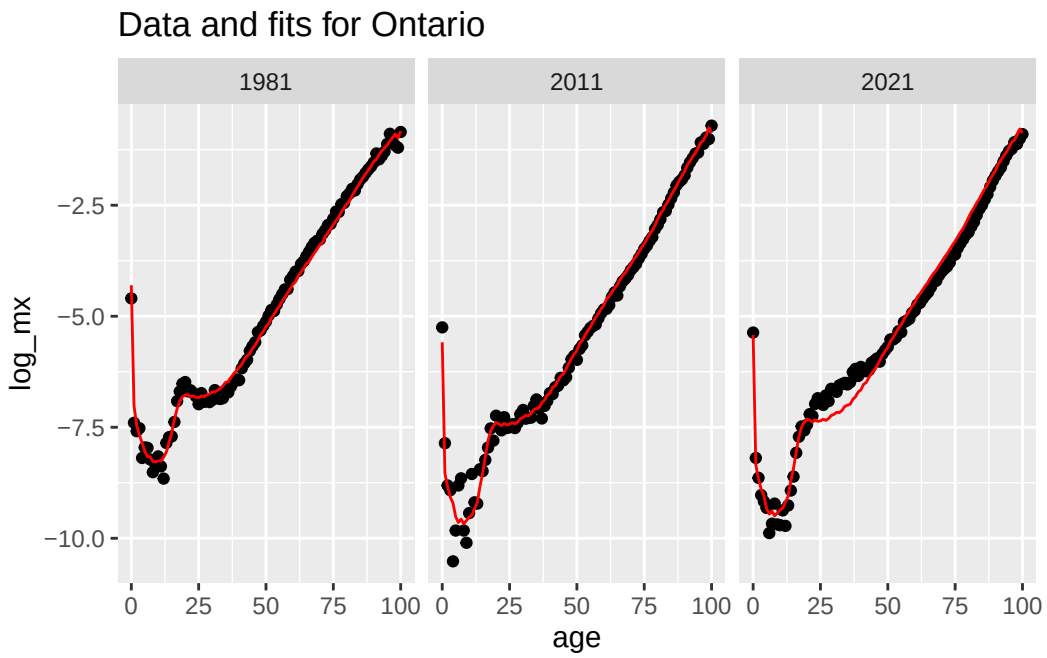
```r
#Original Lee-carter
m_tx <- dm |>
  filter(age < 101) |>
  select(year, age, male) |>
  pivot_wider(names_from = "age", values_from = "male") |>
  select(-year) |>
  as.matrix()

ages <- 0:100
years <- unique(dm$year)
logm_tx <- log(m_tx)
logm_tx[is.infinite(logm_tx)] <- min(logm_tx[!is.infinite(logm_tx)])
ax <- apply(logm_tx, 2, mean)

plot(ax)
```




```r
swept_logm_tx <- sweep(logm_tx, 2, ax)

svd_mx <- svd(swept_logm_tx)

names(svd_mx)
```

```
[1] "d" "u" "v"
```

```
bx <- svd_mx$v[, 1]/sum(svd_mx$v[, 1])
kt <- svd_mx$d[1] * svd_mx$u[, 1] * sum(svd_mx$v[, 1])

lc_age_df <- tibble(age = ages, ax = ax, bx = bx)
lc_time_df <- tibble(year = years, kt = kt)


data_and_res <- dm |>
  filter(age < 101) |>
  mutate(log_mx = log(male)) |>
  left_join(lc_age_df) |>
  left_join(lc_time_df) |>
  mutate(lc_fit = ax + bx*kt)

data_and_res |>
  filter(year %in% c(1981, 2011, 2021)) |>
  ggplot(aes(age, log_mx)) + geom_point() +
  facet_wrap(~year) +
  geom_line(aes(age, lc_fit), color = "red") +
  ggtitle("Data and fits for Ontario")
```



```r
m_tx_1970 <- dm |>
  filter(year >= 1970, age < 101) |>
  select(year, age, male) |>
  pivot_wider(names_from = "age", values_from = "male") |>
  select(-year) |>
  as.matrix()

ages <- 0:100

dm1970 <- dm |>
  filter(year >= 1970, age < 101)

years <- unique(dm1970$year,)
```

```r
logm_tx_1970 <- log(m_tx_1970)
logm_tx_1970[is.infinite(logm_tx_1970)] <- min(logm_tx_1970[!is.infinite(logm_tx_1970)])
ax_1970 <- apply(logm_tx_1970, 2, mean)
```

```r
#demeaning

swept_logm_1970 <- sweep(logm_tx_1970, 2, ax_1970)

svd_1970 <- svd(swept_logm_1970)

bx1970 <- svd_1970$v[, 1]/sum(svd_1970$v[, 1])

kt_1970 <- svd_1970$d[1] * svd_1970$u[, 1] * sum(svd_1970$v[, 1])
```

```r
lc_age_df1970 <- tibble(age = ages, ax_1970 = ax_1970, bx1970 = bx1970)
lc_time_df1970 <- tibble(year = years, kt_1970 = kt_1970)
```

```r
data_and_res <- dm |>
  filter(age < 101) |>
  mutate(logm_tx_1970 = log(male)) |>
  left_join(lc_age_df1970) |>
  left_join(lc_time_df1970) |>
  mutate(lc_fit1970 = ax_1970 + bx1970*kt_1970)

data_and_res |>
  filter(year %in% c(1981, 2011, 2021)) |>
  ggplot(aes(age, logm_tx_1970)) + geom_point() +
```

```
  facet_wrap(~year) +
  geom_line(aes(age, lc_fit1970), color = "red") +
  ggtitle("Data and fits for Ontario")
```