Week 3: Mortality II

SOC6708 ADA

Monica Alexander

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

Decomposition

Let's read in WPP data from the first week week and calculate the age-specific mortality rates:

```
d_male <- read_xlsx(here("data/WPP2024_POP_F01_2_POPULATION_SINGLE_AGE_MALE.xlsx"), skip = 16
d_male$sex <- "Male"
d_male <- d_male |> drop_na(Year)
d_female <- read_xlsx(here("data/WPP2024_POP_F01_3_POPULATION_SINGLE_AGE_FEMALE.xlsx"), skip
d_female$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 <- read_xlsx(here("data/WPP2024_MORT_F01_2_DEATHS_SINGLE_AGE_MALE.xlsx"), skip = 16)
d_male$sex <- "Male"
d_male <- d_male |> drop_na(Year)
```

```
d_female <- read_xlsx(here("data/WPP2024_MORT_F01_3_DEATHS_SINGLE_AGE_FEMALE.xlsx"), skip = 1
d_female$sex <- "Female"
d_female <- d_female |> drop_na(Year)

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

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)
```

Do the decomposition of the difference between Kenya and Canada:

```
asmr |>
  filter(region == "Kenya",year==2023)  |>
  select(sex, age, pop, mx) |>
  rename(pop_kenya = pop, mx_kenya = mx) |>
  left_join(asmr |>
              filter(region == "Canada",year==2023)  |>
              select(sex, age, pop, mx) |>
              rename(pop_can = pop, mx_can = mx) ) |>
  mutate(prop_kenya = pop_kenya/sum(pop_kenya),
         prop_can = pop_can/sum(pop_can)) |>
  mutate(rate_diff = mx_kenya - mx_can,
         prop_diff = prop_kenya - prop_can) |>
  mutate(ave_rate = (mx_kenya+mx_can)/2,
```

```
        ave_prop = (prop_kenya+prop_can)/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         -0.0107          0.00998  -0.000724
```

Check that the difference is actually the difference between the two CDRs

```
asmr |>
  filter(region == "Kenya"|region=="Canada",year==2023) |>
  group_by(region) |>
  summarize(cdr = sum(mx*pop)/sum(pop)) |>
  summarise(diff = cdr[region=="Kenya"] - cdr[region=="Canada"])
```

```
# A tibble: 1 x 1
       diff
      <dbl>
1 -0.000724
```

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?

Mortality models

Read in mortality rates for Ontario. These data come from the Canadian Human Mortality Database.

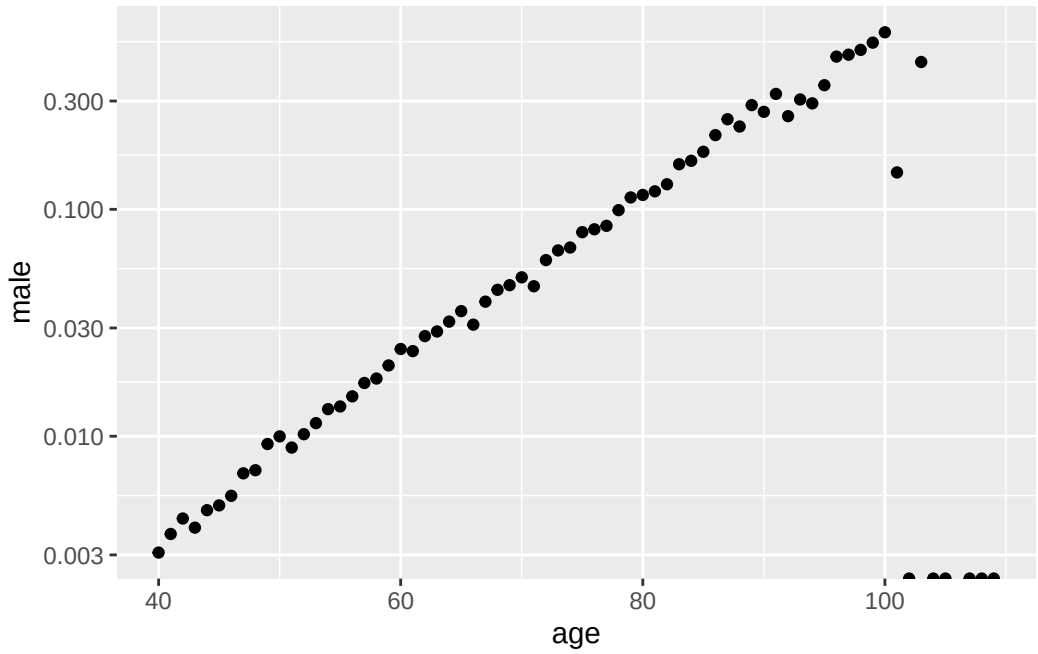
```
dm <- read_table("https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt", skip = 2, col_type
head(dm)
```

```
# 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
```

Gompertz

Let's fit a Gompertz model to Male mortality rates in the year 1950 from age 40. What is the interpretation of the coefficient estimates?

```
# clean up a bit
dm <- dm |>
  clean_names() |>
  mutate(age = as.numeric(age))

df_1950_40 <- dm |>
  filter(year==1950, age>39) |>
  select(age, male)

df_1950_40 |>
  ggplot(aes(age, male)) +
  geom_point()+
  scale_y_log10()
```




```
# remove above 100
df_1950_40 <- df_1950_40 |>
  filter(age<100)

summary(lm(log(male)~age, data = df_1950_40))
```

```
Call:
lm(formula = log(male) ~ age, data = df_1950_40)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26148 -0.05565  0.01617  0.06663  0.17755

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.0255823  0.0531701  -169.7   <2e-16 ***
age          0.0857402  0.0007423   115.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09958 on 58 degrees of freedom
Multiple R-squared:  0.9957,    Adjusted R-squared:  0.9956
F-statistic: 1.334e+04 on 1 and 58 DF,  p-value: < 2.2e-16
```

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.

Lee-Carter

Let's get the Lee-carter model parameters for Ontario. First, get the matrix of age-specific rates:

```
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)
```

log and demean those rates:

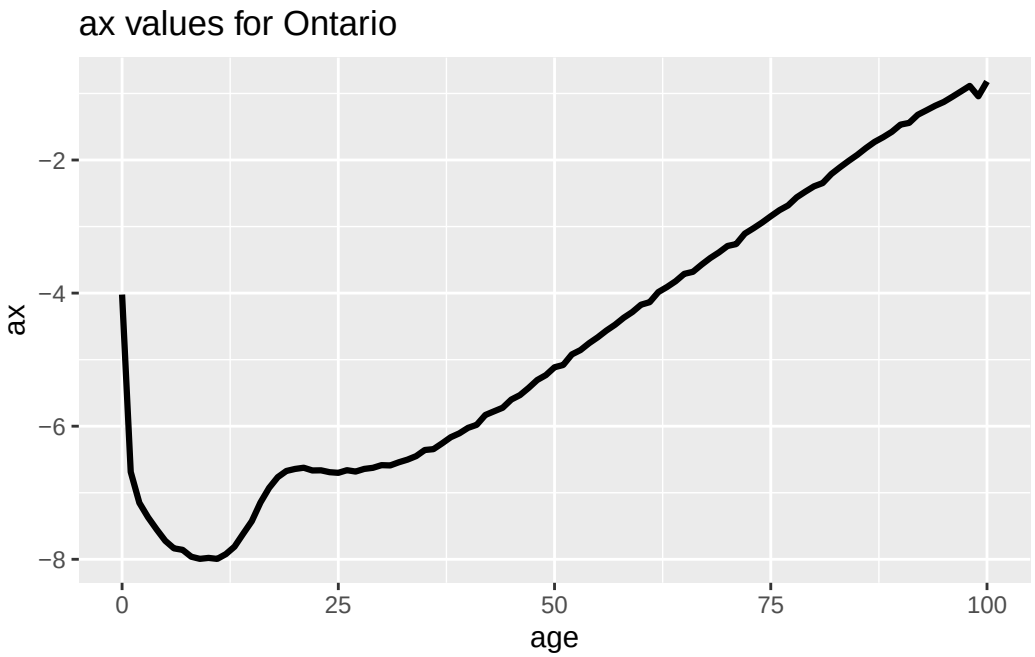
```
logm_tx <- log(m_tx)
logm_tx[is.infinite(logm_tx)] <- min(logm_tx[!is.infinite(logm_tx)])
ax <- apply(logm_tx, 2, mean)
```

Do the SVD

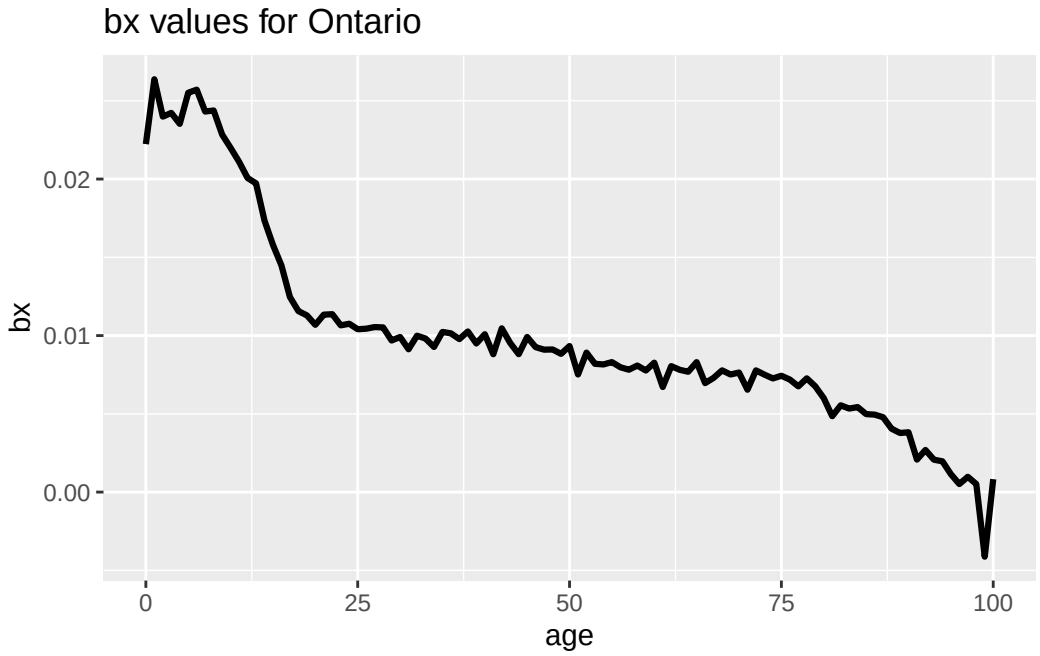
```
swept_logm_tx <- sweep(logm_tx, 2, ax)

svd_mx <- svd(swept_logm_tx)

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

Plots!

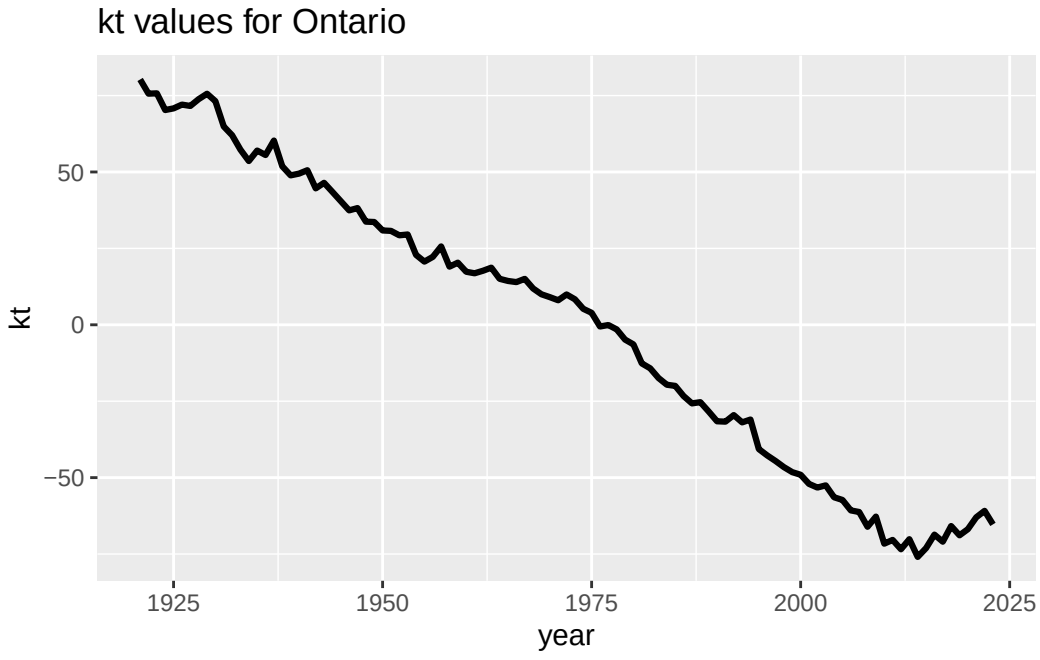
```r
# plot ax

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

ggplot(lc_age_df, aes(age, ax)) +
  geom_line(lwd = 1.1) +
  ggtitle("ax values for Ontario")
```




```r
ggplot(lc_age_df, aes(age, bx)) +
  geom_line(lwd = 1.1) +
  ggtitle("bx values for Ontario")
```

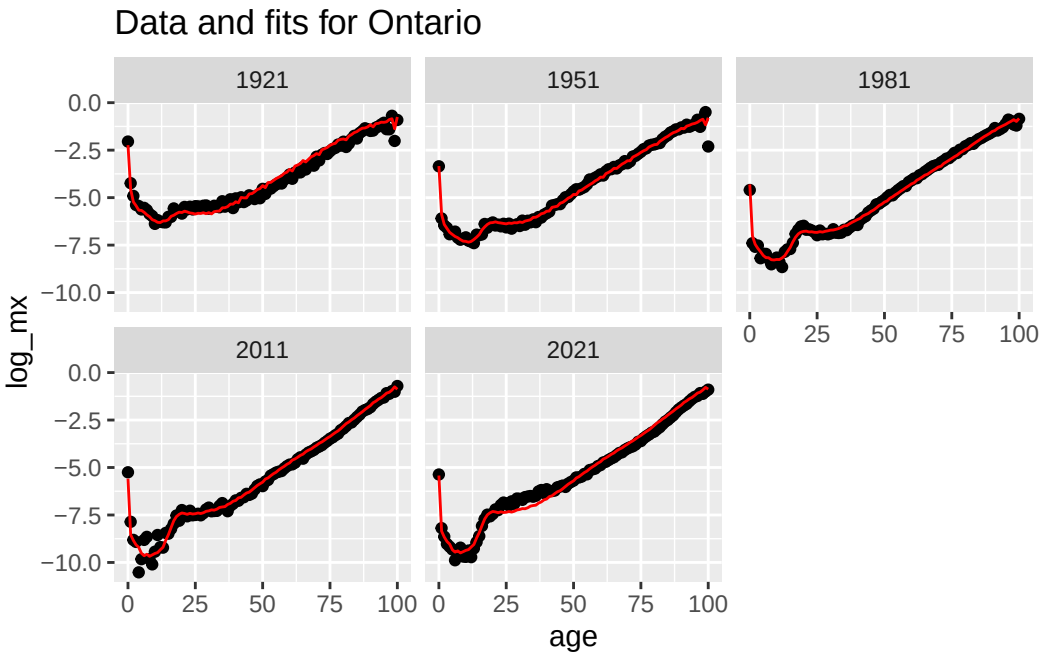



```
ggplot(lc_time_df, aes(year, kt)) +
  geom_line(lwd = 1.1) +
  ggtitle("kt values for Ontario")
```



let's look at the fit for a couple of years

```r
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(1921, 1951, 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")
```



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?