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

```r
d_female <- read_xlsx(here("WPP2024_MORT_F01_3_DEATHS_SINGLE_AGE_FEMALE.xlsx"), skip = 16)
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:

```r
asmr |>
  # get data for kenya
  filter(region == "Kenya",year==2023)  |>
  select(sex, age, pop, mx) |>
  rename(pop_kenya = pop, mx_kenya = mx) |>
  # get data for canada
  left_join(asmr |>
              filter(region == "Canada",year==2023)  |>
              select(sex, age, pop, mx) |>
              rename(pop_can = pop, mx_can = mx) ) |>
  # calculate population proportions
  mutate(prop_kenya = pop_kenya/sum(pop_kenya),
         prop_can = pop_can/sum(pop_can)) |>
```

```
  mutate(rate_diff = mx_kenya - mx_can, # difference in mortality rates
         prop_diff = prop_kenya - prop_can) |> # diference in the pop proportions
  mutate(ave_rate = (mx_kenya+mx_can)/2, # average mortality
         ave_prop = (prop_kenya+prop_can)/2) |> # average pop proportion
  mutate(age_contr = prop_diff*ave_rate, # age contribution
         rate_contr = rate_diff*ave_prop) |> # mortality contribution
  summarize(age_total_contr = sum(age_contr), # summing up over age
            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?

```
asmr |>
  # get data for usa
  filter(region == "United States of America",year==2023)  |>
  select(sex, age, pop, mx) |>
  rename(pop_usa = pop, mx_usa = mx) |>
  # get data for jpn
  left_join(asmr |>
```

```
              filter(region == "Japan",year==2023)  |>
              select(sex, age, pop, mx) |>
              rename(pop_jpn = pop, mx_jpn = mx) ) |>
  # calculate population proportions
  mutate(prop_usa = pop_usa/sum(pop_usa),
         prop_jpn = pop_jpn/sum(pop_jpn)) |>
  mutate(rate_diff = mx_usa - mx_jpn, # difference in mortality rates
         prop_diff = prop_usa - prop_jpn) |> # diference in the pop proportions
  mutate(ave_rate = (mx_usa+mx_jpn)/2, # average mortality
         ave_prop = (prop_usa+prop_jpn)/2) |> # average pop proportion
  mutate(age_contr = prop_diff*ave_rate, # age contribution
         rate_contr = rate_diff*ave_prop) |> # mortality contribution
  summarize(age_total_contr = sum(age_contr), # summing up over age
            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
```

```
asmr |>
  filter(region == "United States of America"|region=="Japan",year==2023) |>
  group_by(region) |>
  summarize(cdr = sum(mx*pop)/sum(pop)) |>
  summarise(diff = cdr[region=="United States of America"] - cdr[region=="Japan"])
```

```
# A tibble: 1 x 1
      diff
     <dbl>
1 -0.00359
```

Mortality models

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

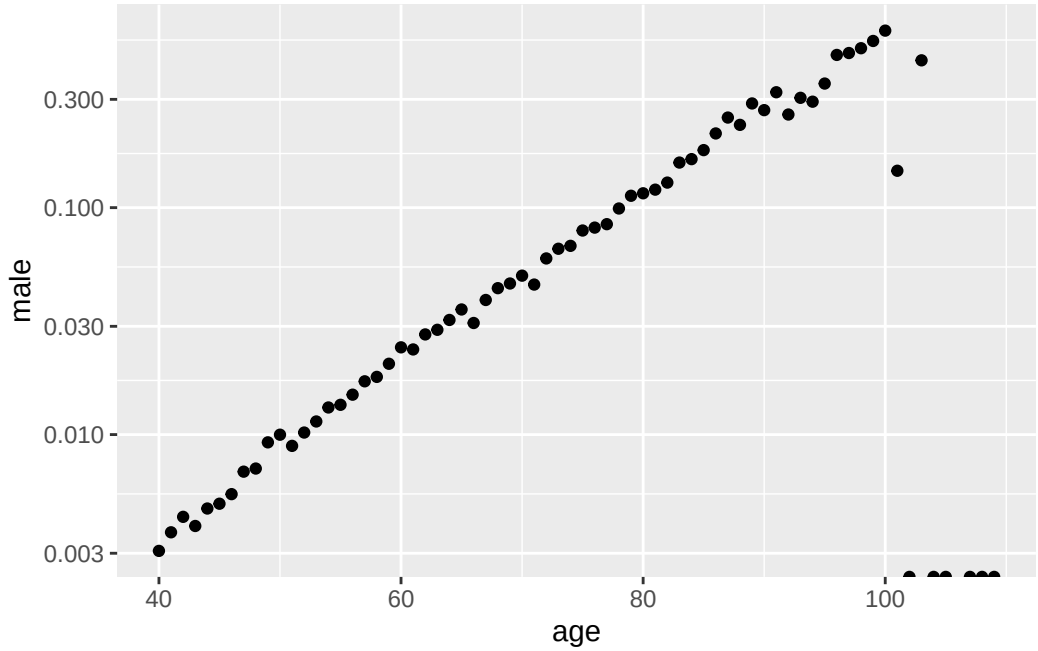
```
dm <- read_table("https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt", skip = 2, col_type
head(dm)
```

```
# 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)

mod <- lm(log(male)~age, data = df_1950_40)

# mortality at age 40 (estimated)
exp(coef(mod)[1])
```

```
 (Intercept)
0.0001202927
```

```
summary(mod)
```

```

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
```

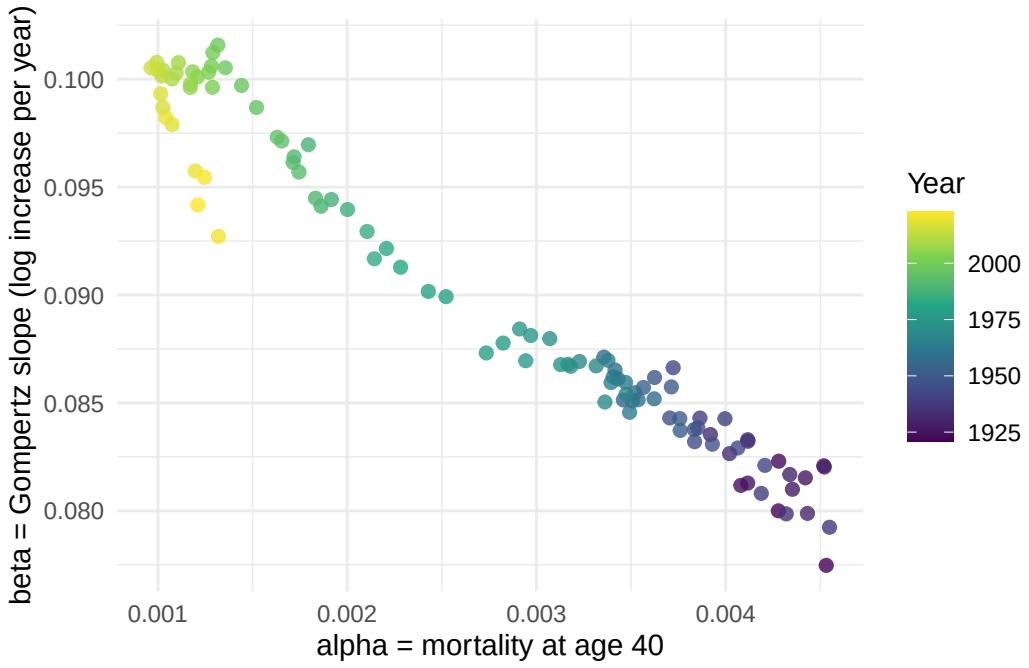
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.0255823  0.0531701  -169.7   <2e-16 ***
age          0.0857402  0.0007423   115.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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.

```r
library(dplyr)
library(tibble)

dm_to_fit <- dm |>
  filter(age>39, age<100) |>
  select(year, age, male)

coefs <- tibble()
years <- sort(unique(dm_to_fit$year))

for (i in seq_along(years)) {
  this_df <- dm_to_fit |>
    filter(year == years[i], is.finite(male), male > 0)

  this_mod <- lm(log(male) ~ I(age - 40), data = this_df)

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

```
}


# Scatter: alpha vs beta, colored by year
ggplot(coefs, aes(x = alpha, y = beta, color = year)) +
  geom_point(size = 2, alpha = 0.8) +
  scale_color_viridis_c() +
  labs(
    x = "alpha = mortality at age 40",
    y = "beta = Gompertz slope (log increase per year)",
    color = "Year"
  ) +
  theme_minimal()
```




```
## mortality level at age 40 declines
## mortality is more concentrated at older ages recent years
```

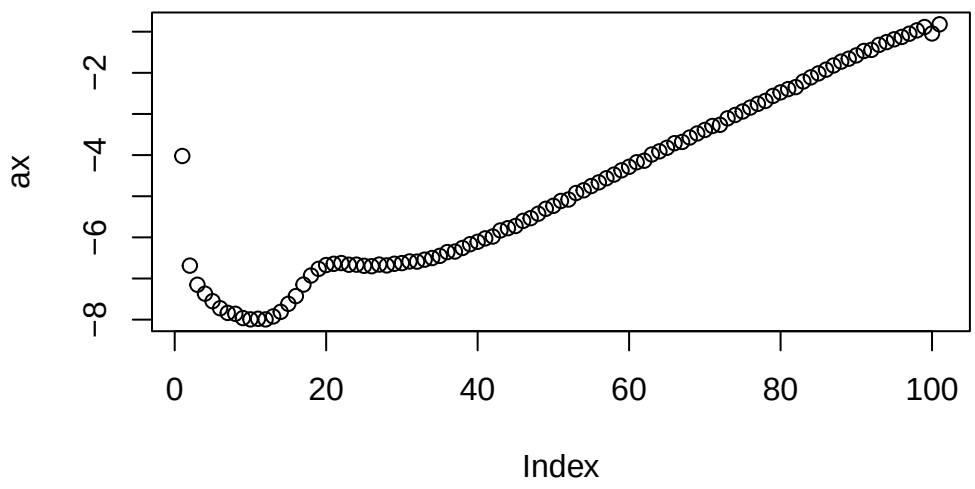
Lee-Carter

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

```r
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:

```r
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)
```



Do the SVD

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

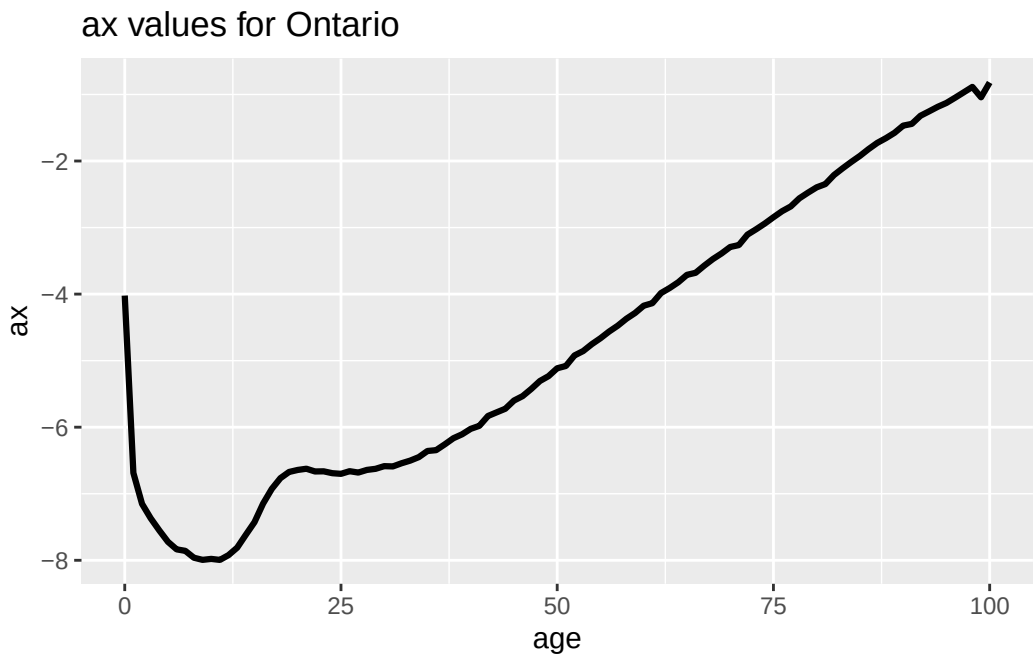
```r
svd_mx <- svd(swept_logm_tx)

names(svd_mx)
```

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

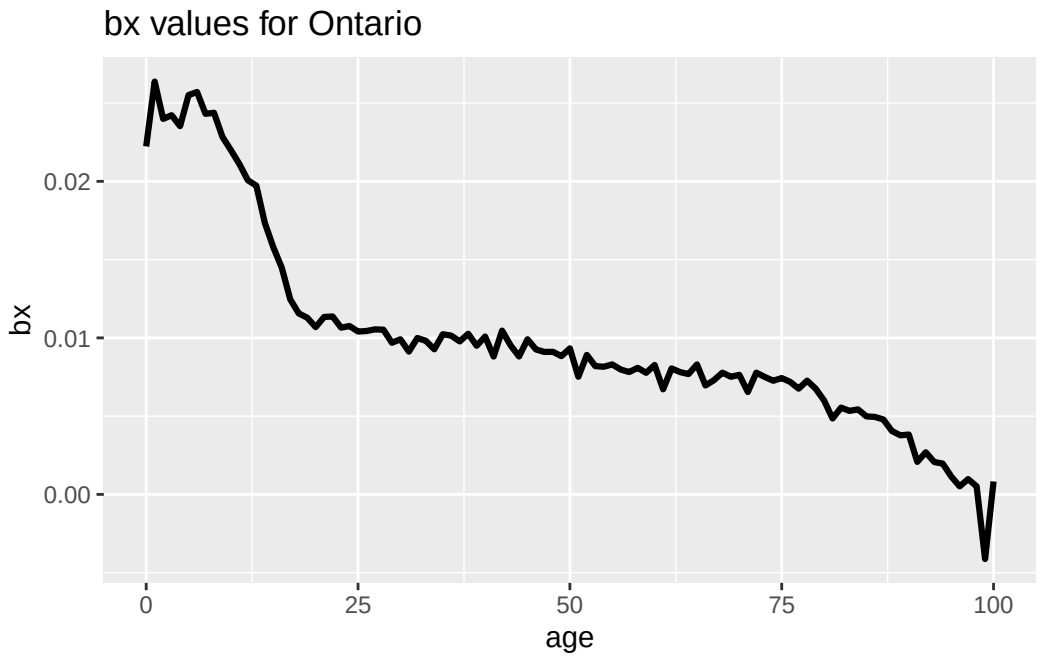
```r
bx <- svd_mx$v[, 1]/sum(svd_mx$v[, 1])
kt <- svd_mx$d[1] * svd_mx$u[, 1] * sum(svd_mx$v[, 1])
```

Plots!

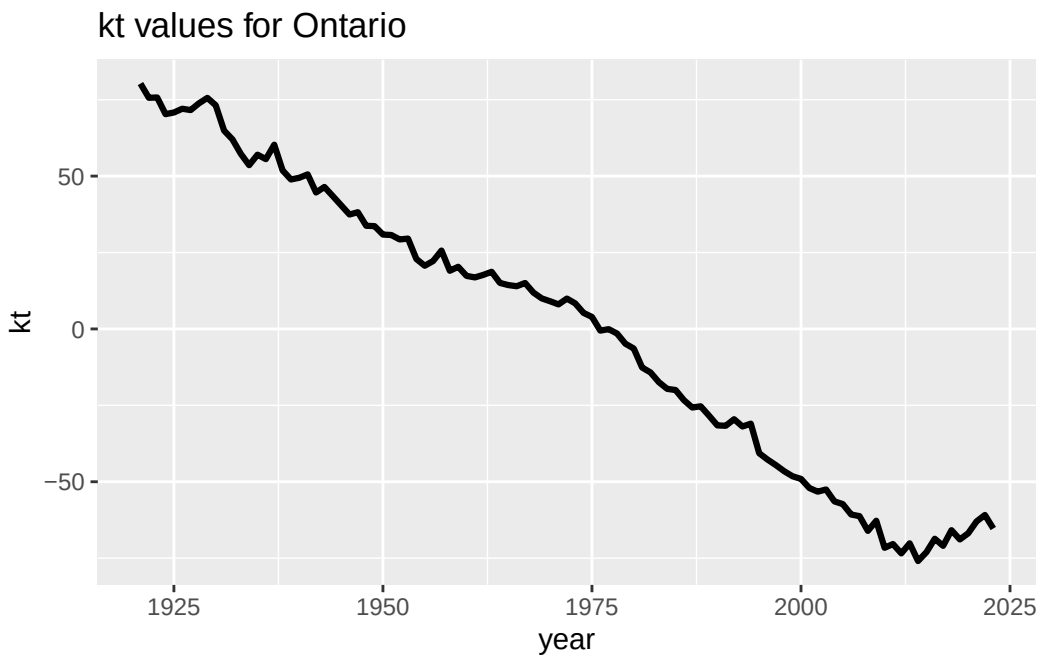
```r
# plot ax

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")
```



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

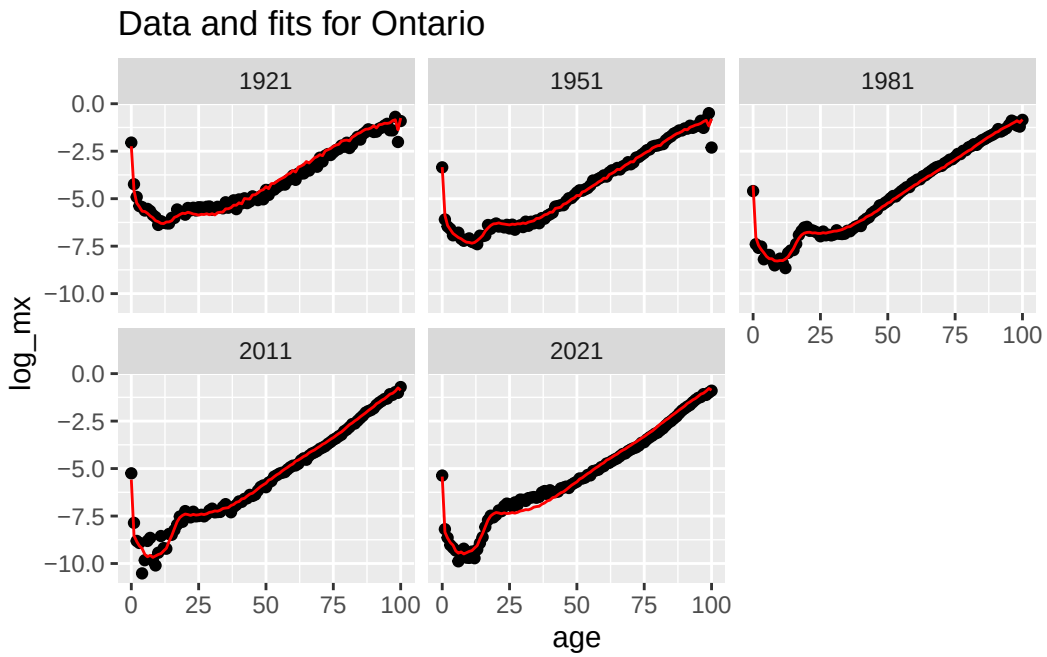



```
ggplot(lc_time_df, aes(year, kt)) +
  geom_line(lwd = 1.1) +
  ggtitle("kt values for Ontario")
```



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

```
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

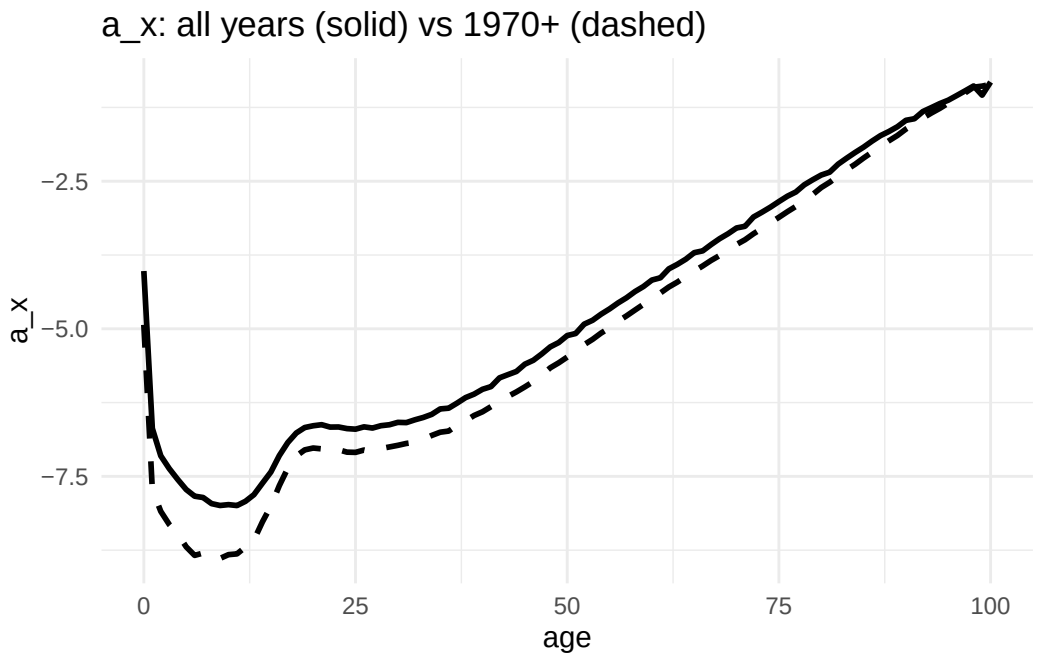
Repeat the lee-carter model fitting exercise but just use mortality rates from 1970. Does this change the estimated rates? Does it do a better or worse job, or does it depend on the year?

```
library(dplyr)
library(tidyr)
library(ggplot2)

fit_lc_1factor <- function(dm_in) {

  dm_sub <- dm_in |>
    filter(age < 101) |>
    select(year, age, male) |>
    arrange(year, age)

  years <- sort(unique(dm_sub$year))
  ages  <- sort(unique(dm_sub$age))

  m_tx <- dm_sub |>
    pivot_wider(names_from = age, values_from = male) |>
    arrange(year) |>
```

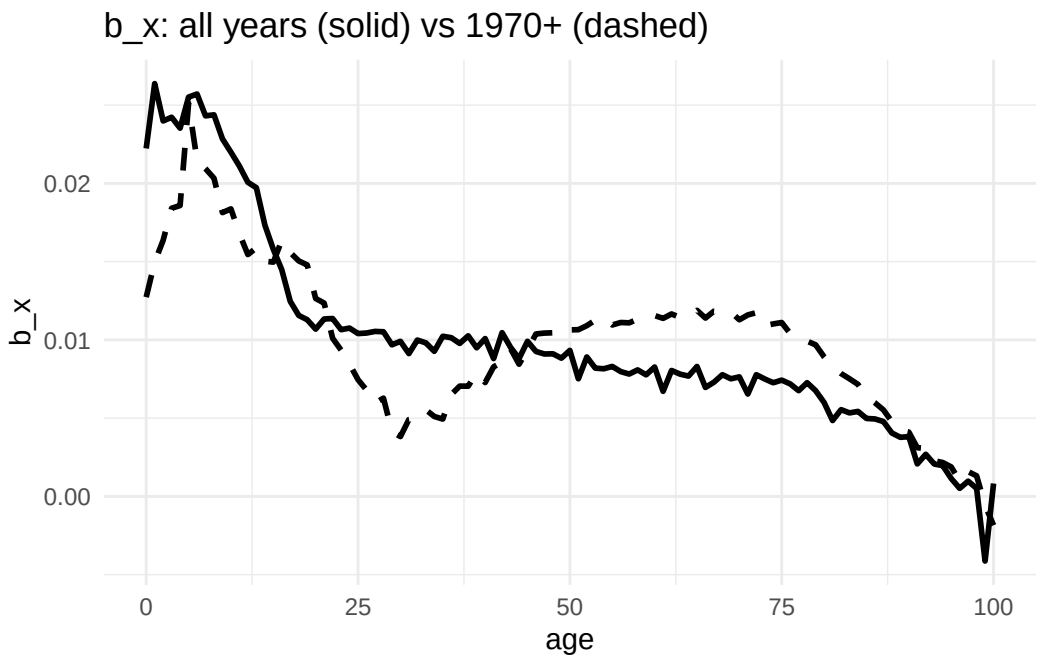
```
    select(-year) |>
    as.matrix()

  logm_tx <- log(m_tx)
  min_finite <- min(logm_tx[is.finite(logm_tx)], na.rm = TRUE)
  logm_tx[!is.finite(logm_tx)] <- min_finite

  ax <- colMeans(logm_tx, na.rm = TRUE)
  Z  <- sweep(logm_tx, 2, ax, "-")

  sv <- svd(Z)
  bx_raw <- sv$v[, 1]
  kt_raw <- sv$d[1] * sv$u[, 1]

  # constraints: sum(bx)=1 and mean(kt)=0
  bx <- bx_raw / sum(bx_raw)
  kt <- kt_raw * sum(bx_raw)

  kt <- kt - mean(kt)
  ax <- ax + bx * mean(kt_raw)

  # optional: force kt to trend downward over time
  if (cor(kt, seq_along(kt)) > 0) {
    bx <- -bx
    kt <- -kt
  }

  list(ax = ax, bx = bx, kt = kt, years = years, ages = ages)
}
```

```
lc_all <- fit_lc_1factor(dm)

lc_70  <- fit_lc_1factor(
  dm |> filter(year >= 1970)
)

age_df <- tibble(
  age = lc_all$ages,
  ax_all = lc_all$ax, bx_all = lc_all$bx,
  ax_70  = lc_70$ax,  bx_70  = lc_70$bx
)
```

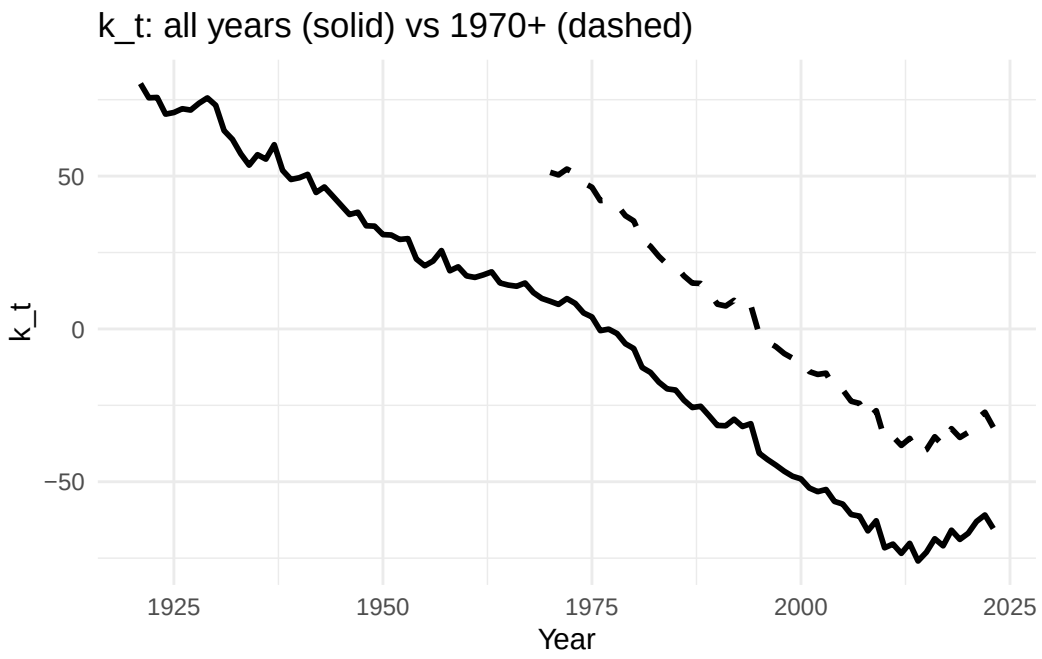
```
ggplot(age_df, aes(age)) +
  geom_line(aes(y = ax_all), linewidth = 1) +
  geom_line(aes(y = ax_70),  linewidth = 1, linetype = 2) +
  labs(title = "a_x: all years (solid) vs 1970+ (dashed)", y = "a_x") +
  theme_minimal()
```




```
ggplot(age_df, aes(age)) +
  geom_line(aes(y = bx_all), linewidth = 1) +
  geom_line(aes(y = bx_70),  linewidth = 1, linetype = 2) +
  labs(title = "b_x: all years (solid) vs 1970+ (dashed)", y = "b_x") +
  theme_minimal()
```



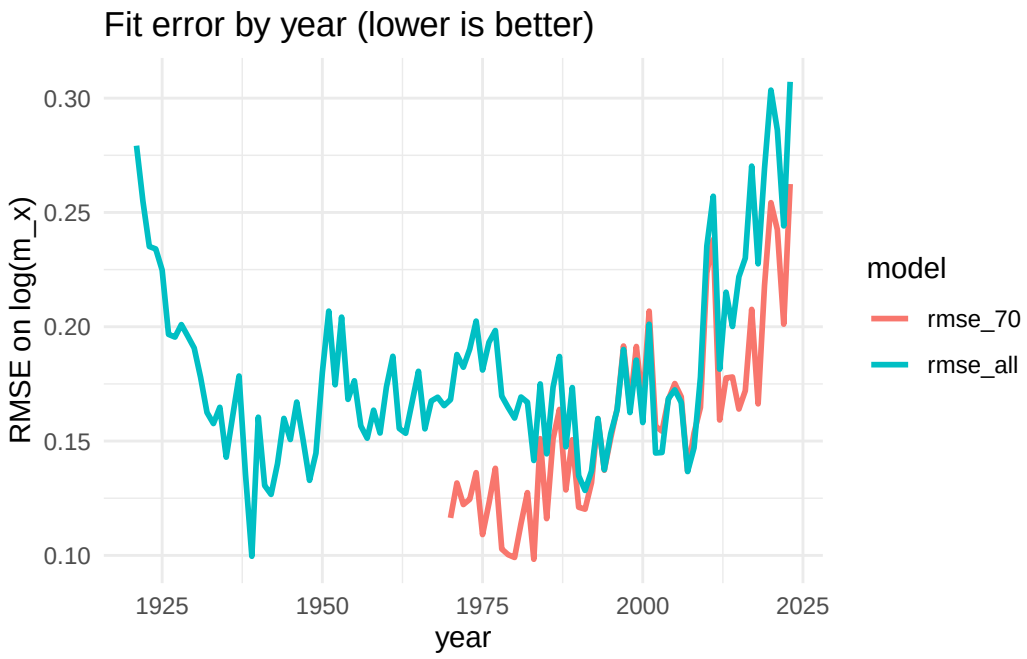

```
kt_df <- tibble(
  year_all = lc_all$years, kt_all = lc_all$kt
)

kt70_df <- tibble(
  year_70 = lc_70$years, kt_70 = lc_70$kt
)

ggplot() +
  geom_line(data = kt_df,  aes(year_all, kt_all), linewidth = 1) +
  geom_line(data = kt70_df, aes(year_70, kt_70), linewidth = 1, linetype = 2) +
  labs(title = "k_t: all years (solid) vs 1970+ (dashed)",
       x = "Year", y = "k_t") +
  theme_minimal()
```




```
# Build a clean evaluation dataset (all years)
eval_df <- dm |>
  filter(age < 101, male > 0) |>
  select(year, age, male) |>
  mutate(log_mx = log(male))

# Create fitted values given an LC fit
add_lc_fit <- function(df, lc, tag) {
  axbx <- tibble(age = lc$ages, ax = lc$ax, bx = lc$bx)
  kt   <- tibble(year = lc$years, kt = lc$kt)

  df |>
    left_join(axbx, by = "age") |>
    left_join(kt,   by = "year") |>
    mutate(!!paste0("fit_", tag) := ax + bx * kt) |>
    select(-ax, -bx, -kt)
}

eval2 <- eval_df |>
  add_lc_fit(lc_all, "all") |>
  add_lc_fit(lc_70,  "70")    # note: pre-1970 years will have NA fit_70 (no kt)

rmse_by_year <- eval2 |>
```

```r
  group_by(year) |>
  summarise(
    rmse_all = sqrt(mean((log_mx - fit_all)^2, na.rm = TRUE)),
    rmse_70  = sqrt(mean((log_mx - fit_70 )^2, na.rm = TRUE)),
    .groups = "drop"
  ) |>
  mutate(diff = rmse_70 - rmse_all)
```

```r
rmse_by_year |>
  pivot_longer(c(rmse_all, rmse_70), names_to = "model", values_to = "rmse") |>
  ggplot(aes(year, rmse, color = model)) +
  geom_line(linewidth = 1.0) +
  theme_minimal() +
  labs(title = "Fit error by year (lower is better)", y = "RMSE on log(m_x)")
```



yes, it changes. It is better.