Homework 04

Ihsan Kahveci

2022-05-16

Contents

1 Questions

1.1 *Q1*

Table 1: Nicaragua female life expectancy at birth and observed gains, 1950-2020

Period Start	e_0	Gain
1950	43.76	2.99
1955	46.75	3.18
1960	49.93	3.31
1965	53.24	3.30
1970	56.54	3.06
1975	59.60	2.89
1980	62.49	3.20
1985	65.69	3.19
1990	68.88	2.28
1995	71.16	2.38
2000	73.54	1.33
2005	74.87	1.51
2010	76.38	1.28
2015	77.66	-

1.1.1 *Q1.a*

Following lecture slides, we can model the expected gain ($\ell_{t+1} - \ell t$) using double-logistic function for a single country, as follows:

$$g(\ell \mid \theta) = \frac{k}{1 + exp(-\frac{2log(9)}{\Delta_2}(\ell - \Delta_1 - 0.5\Delta_2))} - \frac{z - k}{1 + exp(-\frac{2log(9)}{\Delta_4}(\ell - \Delta_1 - \Delta_2 - \Delta_3 - 0.5\Delta_4))}$$

In order to model gains in life expectancy at birth, we use the six parameter double logistic model defined in class. Double-logistic model is harder to converge with numerical optimization because there can be many different set of variables that result in the same DL curve. To make the model converge, I used the starting parameters (Δ_n, k, and z.) of the Markov Chain 1.

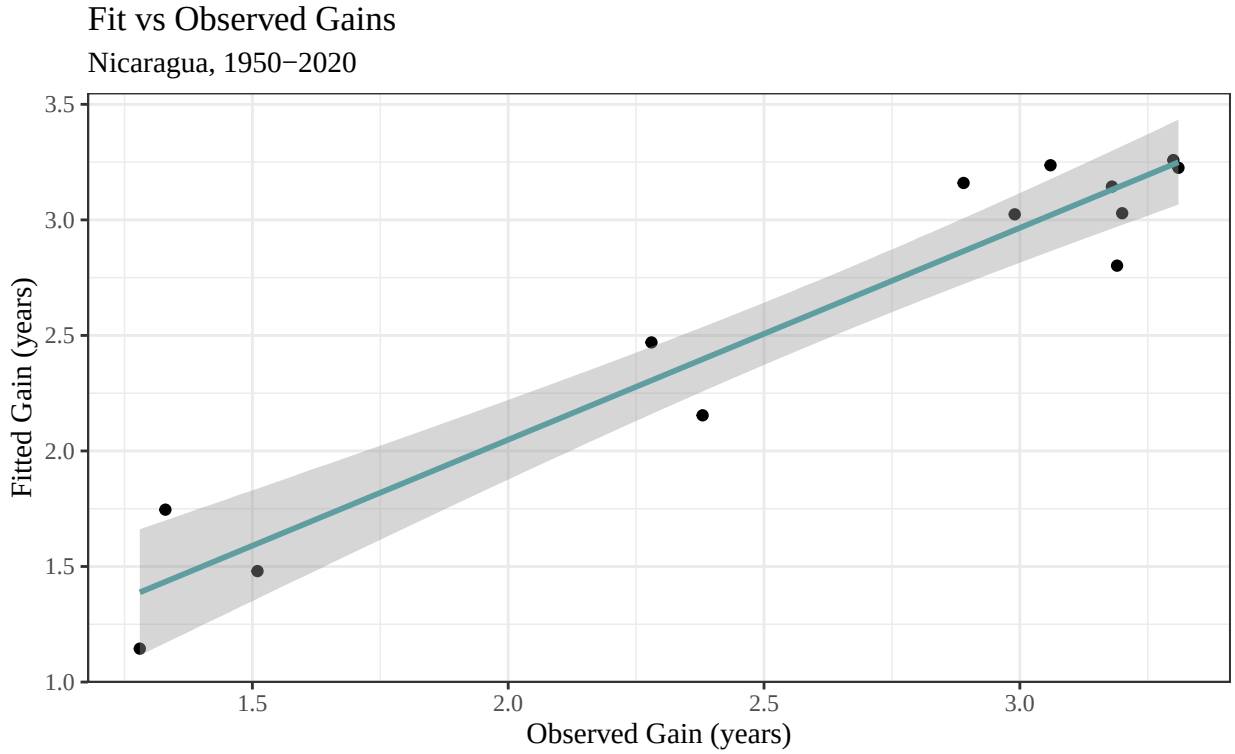
Table 2: Inital Values for Least-Squares Double-Logistic Model

params	values
d1	27.333
d2	39.883
d3	7.132
d4	29.362
k	4.210
z	0.322

With these starting values, we use `optim()` to minimize the least-squares-error, loss of the observed gains vs the fitted gains from the double logistic gain model, which give us the set of optimized parameters $[7.599, 42.019, 18.497, 36.429, 3.722, -7.306]$ with an error variance of 0.574.

1.1.2 *Q1.b*

We can plot use our optimized parameters to get estimates of e_0 gain, and compare them to the observed gains to get a sense of how well the optimization performed:



The model fit considerably well to the observed data, there is a strong correlation (0.96) between fitted and observed values.

1.1.3 *Q1.c*

$$\ell_{2020} = \ell_{2015} + g(\ell_{2015} \mid \theta) + \epsilon$$

$$\sigma^2 = \frac{\sum_{i=1}^{N}(\hat{y_i} - y_i)^2}{N} = 0.5743506$$

Using our observed life expectancy at birth for 2015-2020 and the gain in the same period, we can create an analytic predictive distribution of possible life expectancy at birth for 2020-2025, using the variance σ^2 from our model:

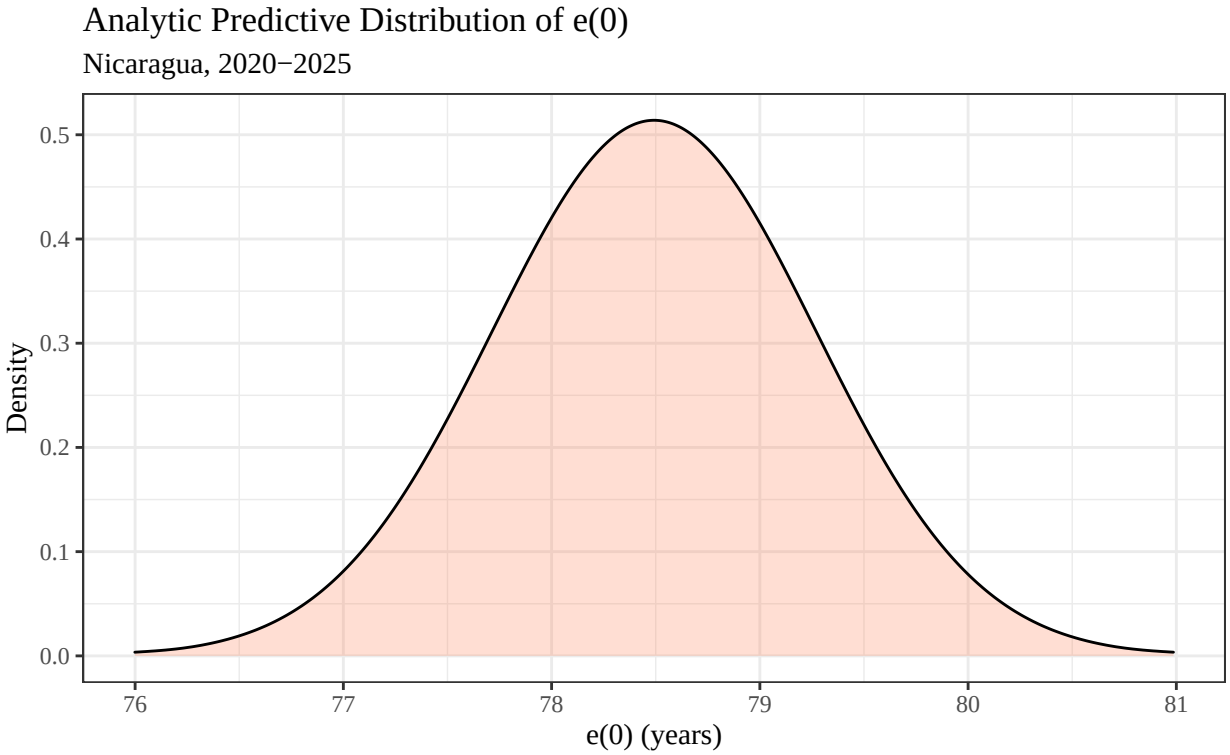


Table 3: Summary of the predictive distribution of 2020-2025 Nicaragua e_0

Mean	Median	2.5% PI	97.5% PI
78.492	78.492	77.007	79.978

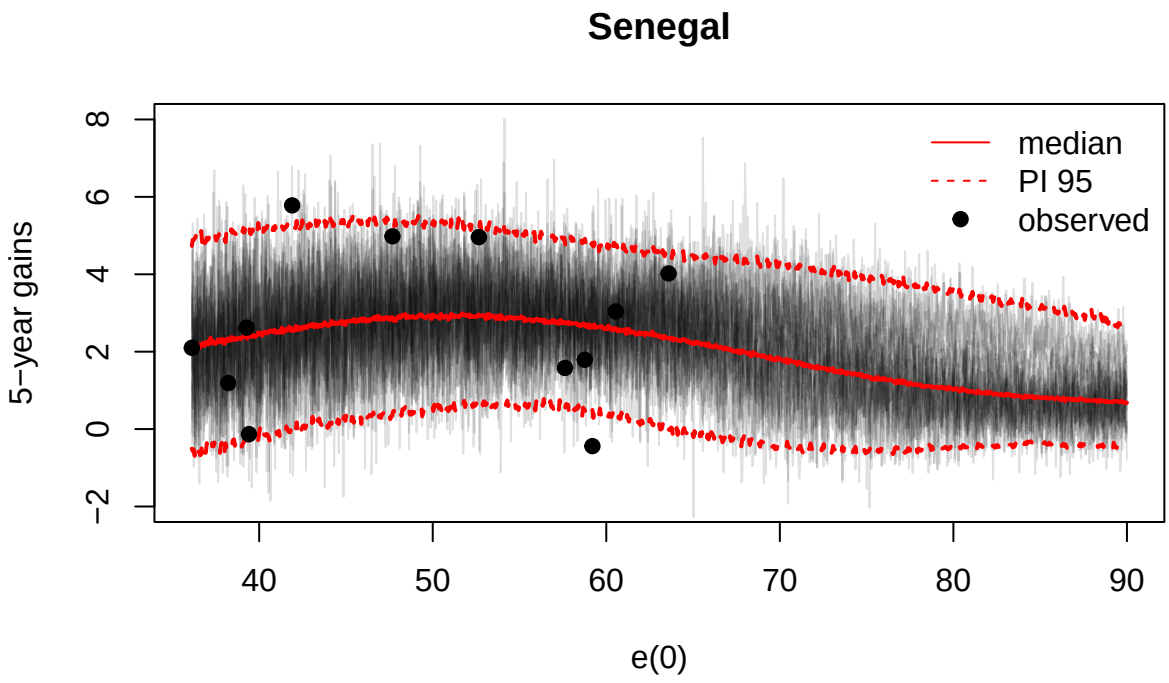
1.2 *Q2*

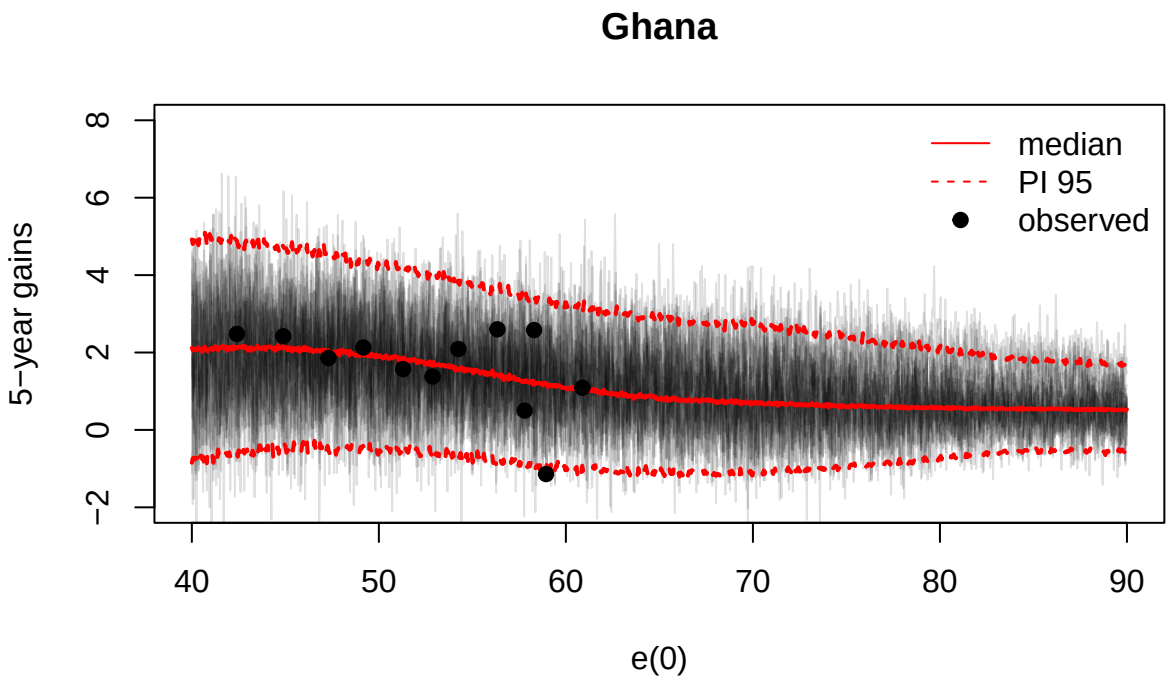
1.2.1 *Q2.a*

Data from a fully converged simulation from a Bayesian model created with *BayesLife* are loaded using the functions `get.e0.mcmc()` and `get.e0.prediction()`. `get.e0.mcmc()` returns an object containing each MCMC chain from the simulation, and `get.e0.prediction()` returns an object containing the summary statistics of the posterior trajectories for the life expectancy created using an input set of MCMC chains.

1.2.2 *Q2.b*

Double logistic curve fits for Senegal and Ghana:





The model fits well to the both countries, as Predictive Intervals include most of the observed points. It seems like it fits Ghana relatively better considering the distribution of observed points. Although the wide predictive intervals prevents to make strong conclusions, the median lines are still insightful. Senegal reaches the peak 5-year gains when life expectancy at birth is approx. 55 years. Whereas Ghana reaches the peak 5-year gains when life expectancy at birth is approx. 45 years. Also, Senegal has a higher expected 5-year gains on average than Ghana until age 80. Then, their trajectory becomes very similar.Also, Senegal also has a wider probability interval in later ages.

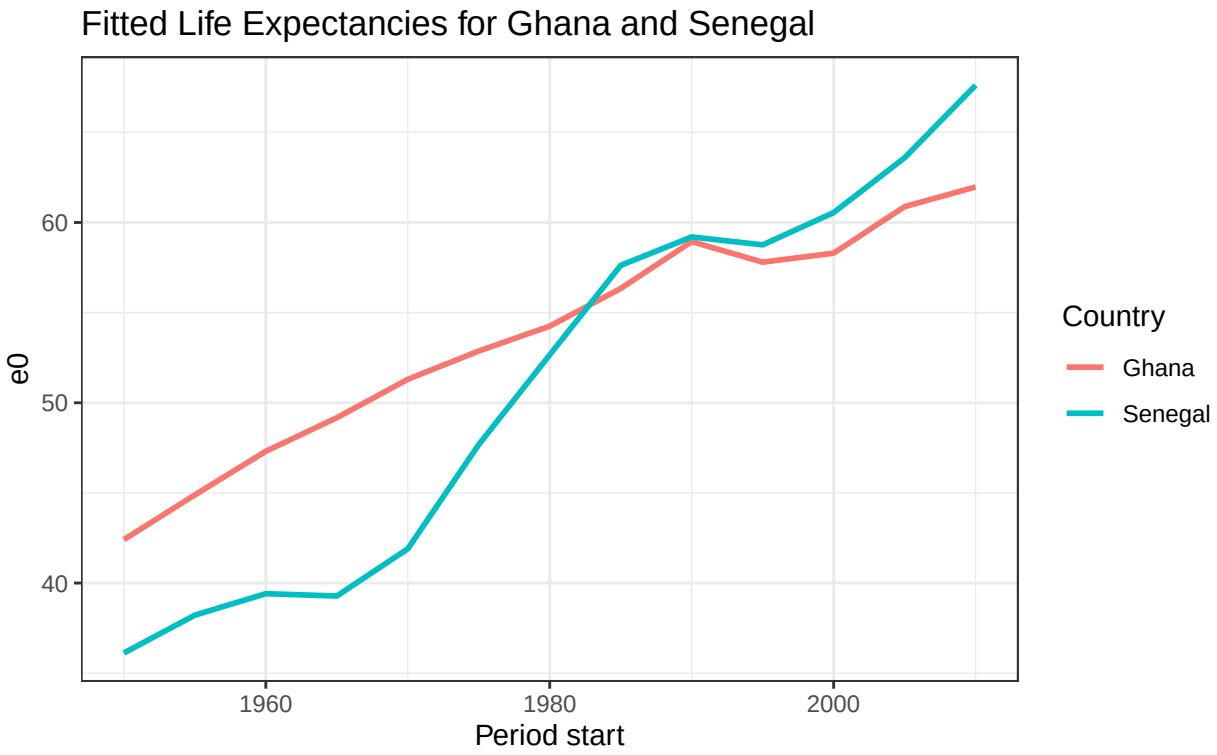
1.2.3 *Q2.c*

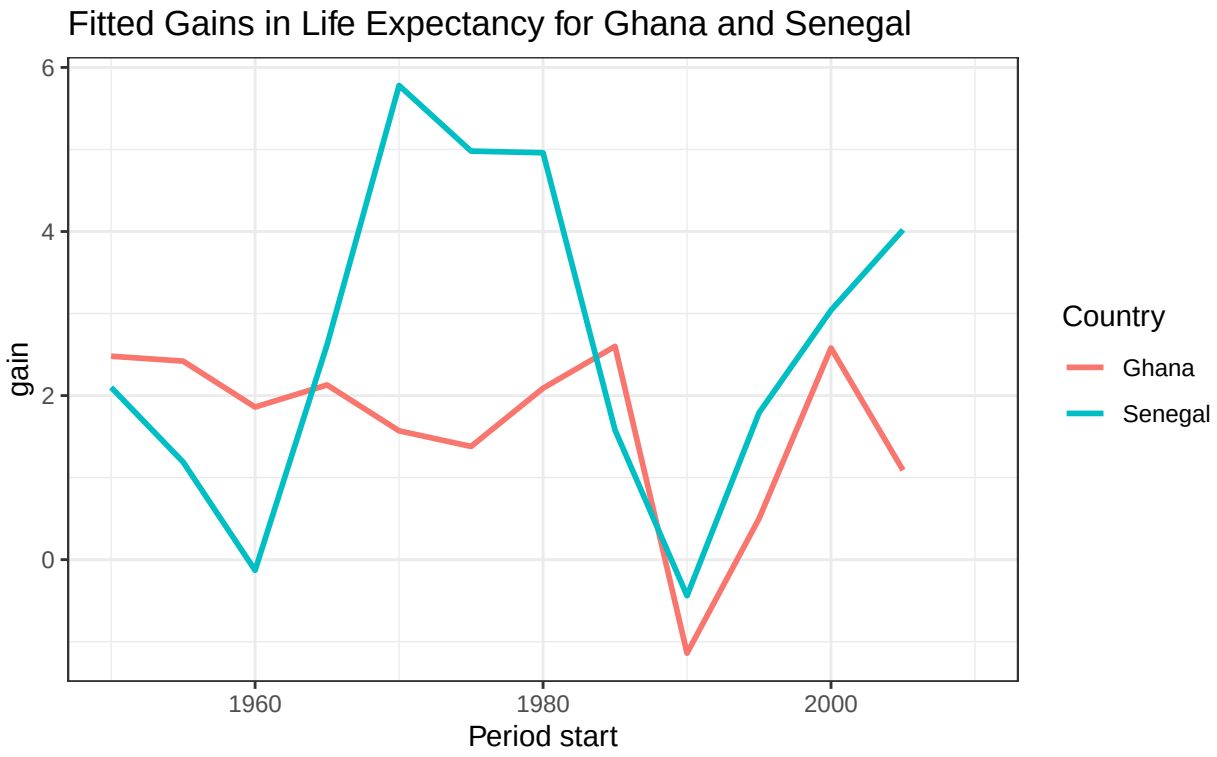
Using `e0.trajectories.table` function, one can easily tabulate the tabulate the posterior distribution of trajectories of the life expectancy for list of countries. Note that years before 2018 doesn't have any uncertainty, as they are observed not predicted. The double-logistic plots indicate that Senegal is experiencing faster increases in life expectancy than Ghana. We can confirm this by examining the reported life expectancies for both countries from 1950 to 2015:

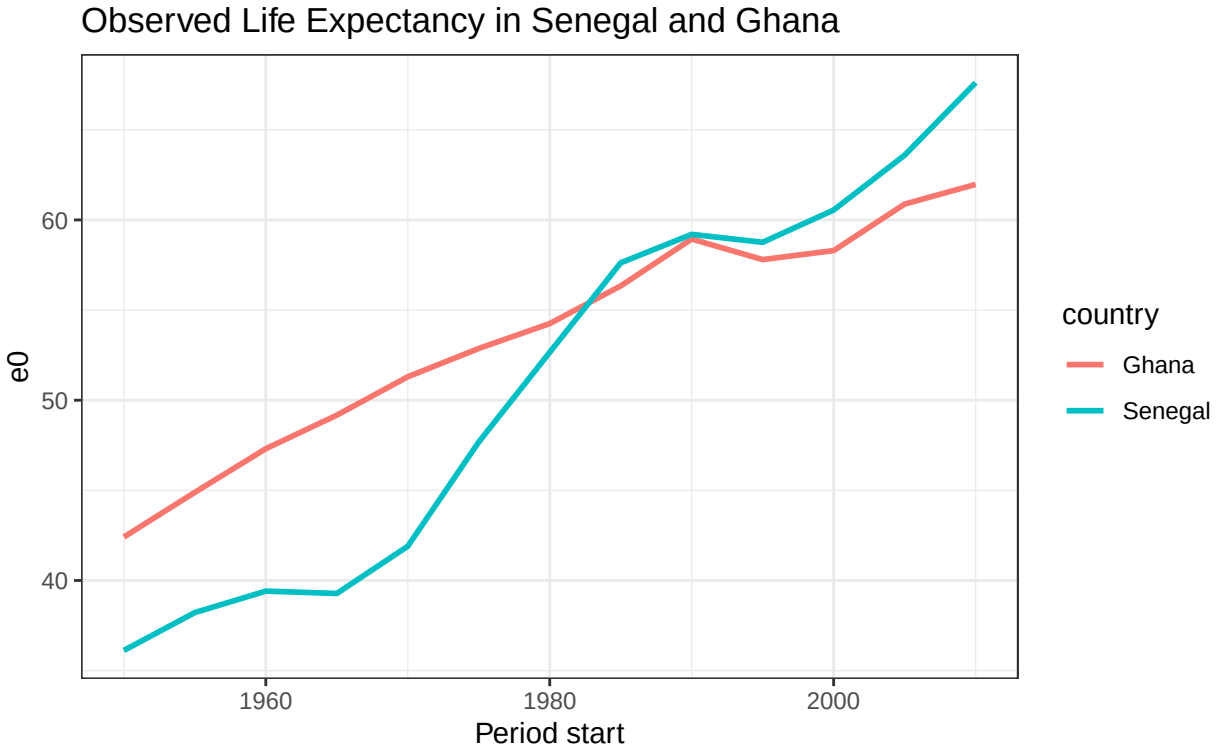
Table 4: Female e_0 for Senegal and Ghana, 1950-2015

Period start	Senegal	Ghana
1950	36.12	42.41
1955	38.22	44.89
1960	39.41	47.31
1965	39.28	49.17
1970	41.90	51.30
1975	47.68	52.87
1980	52.66	54.25
1985	57.62	56.34
1990	59.20	58.94
1995	58.76	57.80
2000	60.55	58.30
2005	63.59	60.88
2010	67.61	61.97

Again, it appears Senegal has a faster increase. We can verify this by calculating the mean gain over time for each country:







Because the existing simulations based on WPP2015 data, I extracted empirical e_0 values from `wpp2015` package. The above plot shows that fitted and observed values are mostly similiar; and Senegal shows a faster increase emprically too.

1.2.4 *Q2.d*

```
## [1] 1000
```

```
## [1] 0.894
```

Assuming gains in life expectancy are independent between countries, we can find the probability of Senegal having either a higher or lower average life expectancy at birth than Ghana in each future period by calculating the number of times Senegal has a higher e_0 than Ghana across all simulations:

$$P(S > G) = \frac{\|S > G\|}{N}$$

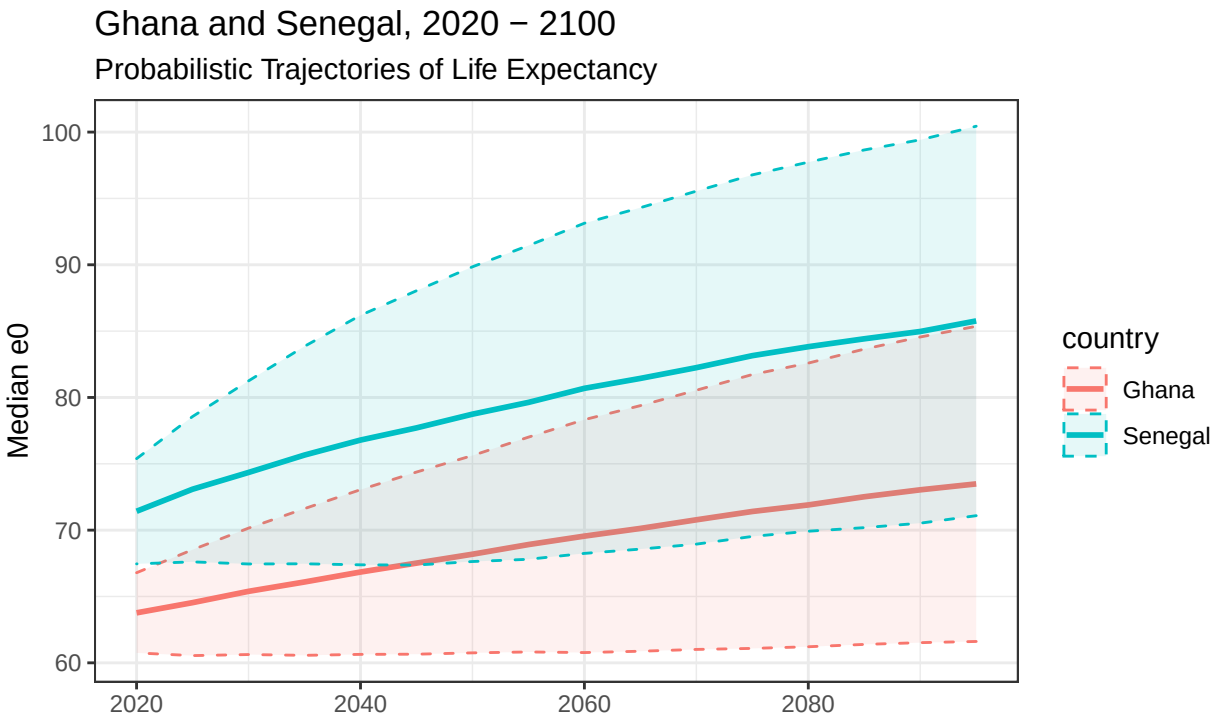
Table 5: Senegal and Ghana Median e_0 Projections in 2020-2025

Period Start	Senegal	Ghana	$P(S > G)$
2020	71.4	63.8	0.997

For all 1000 simulations:
- Probability of Senegal has higher e_0 than Ghana in 2020: 0.997
- Probability of Senegal has higher e_0 in all 16 periods: 0.894

The plot below shows the probability densities of 2020 simulations for Ghana and Senegal. It is clear that the probability of Ghana has higher e_0 than Senegal is very low. (shaded area)

We can also observe the future projections of Senegal and Ghana by plotting them over the years. Confidence Intervals calculated using the SD of the simulations for each year:



2 Appendix

```
# Prep work ---------------------------------------------------------------

# Load libraries
library(tidyverse)
library(bayesLife)
library(bayesPop)

# Data
data("e0F", package = "wpp2019")
e0_sim_dir <- "data/e0/sim03092016"
tfr_sim_dir <- "data/tfr/sim01192018"
pop_sim_dir <- "data/pop/sim05172020"
mig_file <- "data/WPP2019_Period_Indicators_Medium.csv"
can_mig_num_file <- "data/statcan_migration.csv"

e0f_all <- e0F %>%
  select(-country_code, -last.observed) %>%
  pivot_longer(
    cols = -name,
    names_to = "year",
    values_to = "e0",
    names_pattern = "(.*)-"
  ) %>%
  mutate(year = as.integer(year))

# Control randomness
set.seed(57)


# Question 1 --------------------------------------------------------------

e0f_nica <- e0f_all %>%
  filter(name == "Nicaragua") %>%
  select(-name) %>%
  mutate(gain = lead(e0) - e0)

knitr::kable(
  e0f_nica,
  booktabs = TRUE, digits = 2, eval = FALSE,
  col.names = c("Period Start", "$e_0$", "Gain"),
  caption = "Nicaragua female life expectancy at birth and observed gains, 1950-2020"
)


# Question 1a -------------------------------------------------------------
# double-logistic function
# parameters: 4 delta, k, z
dl_gain <- function(l, theta) {

  d1 <- theta[["d1"]]
  d2 <- theta[["d2"]]
```

```
  d3 <- theta[["d3"]]
  d4 <- theta[["d4"]]
  k <- theta[["k"]]
  z <- theta[["z"]]

  (k       / (1 + exp( (-2*log(9) / d2) * (l - d1 - .5*d2         )))) +
  ((z - k) / (1 + exp( (-2*log(9) / d4) * (l - d1 - d2 -d3 - .5*d4))))

}

# least squares error function
# takes a function finds the LS between fitted and observed values
ls_err <- function(func, data, obs_vals) {

  function(params) {
    fit_vals <- func(data, params)
    sum((fit_vals - obs_vals)^2)
  }
}

e0_sim_mcmc<- get.e0.mcmc(e0_sim_dir)
mcmc1 = e0_sim_mcmc$mcmc.list$`1`
#e0.DLcurve.plot(e0_sim_mcmc, country = "Nicaragua")
starting_params <- c(mcmc1$Triangle.ini, mcmc1$k.ini, mcmc1$z.ini)
names(starting_params) <- c("d1", "d2", "d3", "d4", "k", "z")

tibble(params = names(starting_params), values = starting_params) %>%
  knitr::kable(booktabs = TRUE, digits = 3, eval = FALSE,
               caption = "Inital Values for Least-Squares Double-Logistic Model")
# removing last row because it has no gain value
opt_input <- e0f_nica %>% slice(-n())
loss_func <- ls_err(dl_gain, opt_input$e0, opt_input$gain)

opt_result <- optim(starting_params, loss_func)

opt_params <- opt_result$par
opt_err_var <- opt_result$value


# Question 1b ------------------------------------------------------------------

e0f_nica_gains <- e0f_nica %>%
  rename(obs_gain = gain) %>%
  mutate(fit_gain = dl_gain(e0, opt_params))

cor_nica = cor(e0f_nica_gains$obs_gain[1:13], e0f_nica_gains$fit_gain[1:13])
e0f_nica_gains %>%
  drop_na() %>%
  ggplot(aes(x = obs_gain, y = fit_gain)) +
  geom_point() +
  geom_smooth(formula=y~x, method = "lm", color = "cadetblue") +
  theme_bw() +
  theme(text = element_text(family = "serif")) +
```

```
  labs(
    title = "Fit vs Observed Gains",
    subtitle = "Nicaragua, 1950-2020",
    x = "Observed Gain (years)",
    y = "Fitted Gain (years)"
  )


# Question 1c ------------------------------------------------------------

nica_e0_2020_mean <- e0f_nica_gains %>%
  filter(year == 2015) %>%
  select(e0, fit_gain) %>%
  rowSums()

nica_e0_2020_sd <- sqrt(opt_err_var)

nica_e0_2020_dist <- qnorm(
  seq(.0005, .9995, .001),
  mean = nica_e0_2020_mean,
  sd = nica_e0_2020_sd
)

ggplot(enframe(nica_e0_2020_dist), aes(x = value)) +
  geom_density(fill = "coral", alpha = .25) +
  theme_bw() +
  theme(text = element_text(family = "serif")) +
  labs(
    title = "Analytic Predictive Distribution of e(0)",
    subtitle = "Nicaragua, 2020-2025",
    x = "e(0) (years)",
    y = "Density"
  )

nica_e0_2020_tbl <- tibble(
  Mean = nica_e0_2020_mean,
  Median = median(nica_e0_2020_dist),
  `2.5% PI` = Mean - 1.96 * nica_e0_2020_sd,
  `97.5% PI` = Mean + 1.96 * nica_e0_2020_sd
)

knitr::kable(
  nica_e0_2020_tbl,
  booktabs = TRUE, digits = 3, eval = FALSE,
  caption = "Summary of the predictive distribution of 2020-2025 Nicaragua $e_0$"
)


# Question 2a ------------------------------------------------------------

e0_sim_mcmc <- get.e0.mcmc(e0_sim_dir)
e0_sim_pred <- get.e0.prediction(e0_sim_dir)
```

```r
# Question 2b ------------------------------------------------------------------

e0.DLcurve.plot(e0_sim_mcmc, "Senegal",
  predictive.distr = TRUE, pi = c(95), ylim = c(-2, 8))

e0.DLcurve.plot(e0_sim_mcmc, "Ghana",
  predictive.distr = TRUE, pi = c(95), ylim = c(-2, 8))


# Question 2c ------------------------------------------------------------------

e0_traj_tbl <- list(Senegal = "Senegal", Ghana = "Ghana") %>%
  purrr::map(~e0.trajectories.table(e0_sim_pred, country = .x)) %>%
  purrr::map_dfr(~as_tibble(.x, rownames = "year"), .id = "country") %>%
  mutate(year = as.integer(year) - 3) %>%
  filter(year < 2015) %>%
  select(country, `Period start` = year, median) %>%
  pivot_wider(names_from = country, values_from = median)

e0_traj_mean_tbl <- e0_traj_tbl %>%
  pivot_longer(-`Period start`, names_to = "Country", values_to = "e0") %>%
  group_by(Country) %>%
  mutate(gain = lead(e0) - e0)


  # summarise(`Mean Gain` = mean(gain, na.rm = TRUE),
  #           SD = sd(gain, na.rm = TRUE)) %>%
  # mutate(`5% CI` = `Mean Gain` - 1.96 * SD,
  #        `95% CI` = `Mean Gain` + 1.96 * SD) %>%
  # arrange(desc(`Mean Gain`))

knitr::kable(
  e0_traj_tbl,
  booktabs = TRUE, digits = 2, eval = FALSE,
  caption = "Female $e_0$ for Senegal and Ghana, 1950-2015"
)
e0_traj_mean_tbl %>%
  ggplot(aes(x=`Period start`, y=e0, color= Country)) +
  geom_line(size=1) +
  theme_bw() +
  ggtitle("Fitted Life Expectancies for Ghana and Senegal")
e0_traj_mean_tbl %>%
  ggplot(aes(x=`Period start`, y=gain, color= Country)) +
  geom_line(size=1) +
  theme_bw() +
  ggtitle("Fitted Gains in Life Expectancy for Ghana and Senegal")

data("e0F", package = "wpp2015")

e0f_2015 <- e0F %>%
  select(-country_code, -last.observed) %>%
  pivot_longer(
    cols = -country,
```

```
    names_to = "year",
    values_to = "e0",
    names_pattern = "(.*)-"
  ) %>%
  mutate(year = as.integer(year)) %>%
  filter(country %in% c("Senegal", "Ghana"))

e0f_2015 %>%
  mutate(`Period start` = year) %>%
  ggplot(aes(x=`Period start`, y=e0, color= country)) +
  geom_line(size=1) +
  theme_bw() +
  ggtitle("Observed Life Expectancy in Senegal and Ghana")

# Question 2d ------------------------------------------------------------
# tidying the life expectancy projections
sen = get.e0.trajectories(e0_sim_pred, "Senegal")[-c(1,2),]

sen_df = as_tibble(sen) %>%
  mutate(country = "Senegal", year = seq(2020,2095,5), .before=1)

gha = get.e0.trajectories(e0_sim_pred, "Ghana")[-c(1,2),]

gha_df =   as_tibble(gha) %>%
  mutate(country = "Ghana", year = seq(2020,2095,5), .before=1)


sen_gha  = bind_rows(sen_df, gha_df) %>%
  pivot_longer(-c(country, year), names_to = "sim", values_to = "e0")

sen_gha_sum =  sen_gha %>%
  group_by(country, year) %>%
  summarise(median_e0 = median(e0),
            low = median_e0 - 1.96*sd(e0),
            high = median_e0 + 1.96*sd(e0))

sen_gha_probs = sen_gha %>%
  pivot_wider(id_cols = c(sim, year),  names_from = country, values_from = e0) %>%
  mutate(greater = ifelse(Senegal > Ghana, 1, 0)) %>%
  group_by(year) %>%
  summarise(Senegal = median(Senegal), Ghana = median(Ghana), prob = mean(greater))


probs = c()
for (i in 1:ncol(sen)){
  out = sen[,i] > gha[,i]
  n_true = length(out[out==TRUE])
  prob = ifelse(n_true == nrow(sen), 1, 0)
  probs = c(probs, prob)
}
probs %>% length()
mean(probs)
sen_gha_probs %>%
```

```
  filter(year == 2020) %>%
#  pivot_wider(id_cols = year, names_from = country, values_from = median_e0) %>%
  knitr::kable(
  booktabs = TRUE, digits = c(0,1,1,3), eval = FALSE,
  col.names = c("Period Start", "Senegal", "Ghana", "$P(S>G)$"),
  caption = "Senegal and Ghana Median $e_0$ Projections in 2020-2025")
sen_gha %>%
  filter(year == 2020) %>%
  ggplot(aes(x = e0, fill = country)) +
  geom_density(alpha = 0.3) +
  ggtitle("Predictive Distributions for Life Expectancy in 2020", subtitle = "Ghana vs. Senegal") +
  theme_bw()
sen_gha_sum %>%
  ggplot(aes(x=year, y=median_e0, color = country,fill=country)) +
  geom_line(size = 1) +
  geom_ribbon(aes(ymin = low , ymax = high), alpha = 0.1, linetype = "dashed" ) +
  ggtitle("Ghana and Senegal, 2020 - 2100", subtitle = "Probabilistic Trajectories of Life Expectancy")
  labs(x= " ", y = "Median e0") +
  theme_bw()
```