Homework 06

Spencer Pease

5/18/2020

Questions

Q1

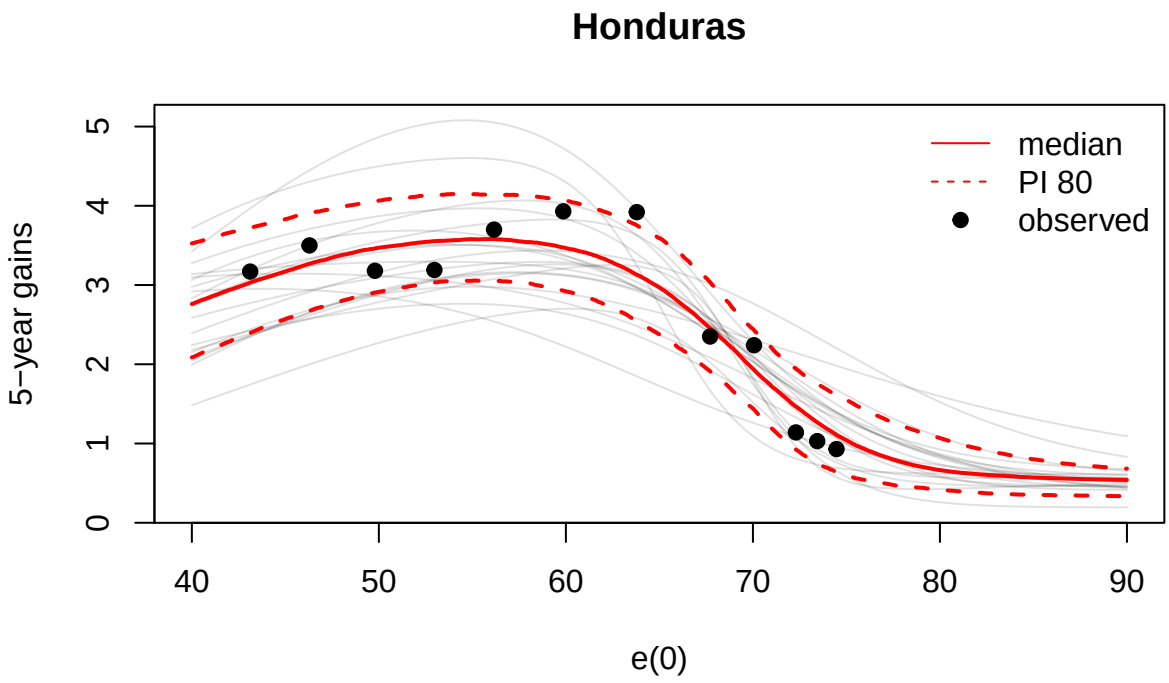
Table 1: Honduras female life expectancy at birth and observed gains, 1950-2020

Period start	e_0	Gain
1950	43.12	3.17
1955	46.29	3.50
1960	49.79	3.18
1965	52.97	3.19
1970	56.16	3.70
1975	59.86	3.93
1980	63.79	3.92
1985	67.71	2.35
1990	70.06	2.24
1995	72.30	1.35
2000	73.65	1.36
2005	75.01	1.30
2010	76.31	0.97
2015	77.28	-

Q1.a

In order to model gains in life expectancy at birth, we use the six parameter double logistic model defined in class.

In order to choose appropriate starting parameters, we look at the double-logistic curve fit for Honduras, and pick out the appropriate deltas, k, and z.

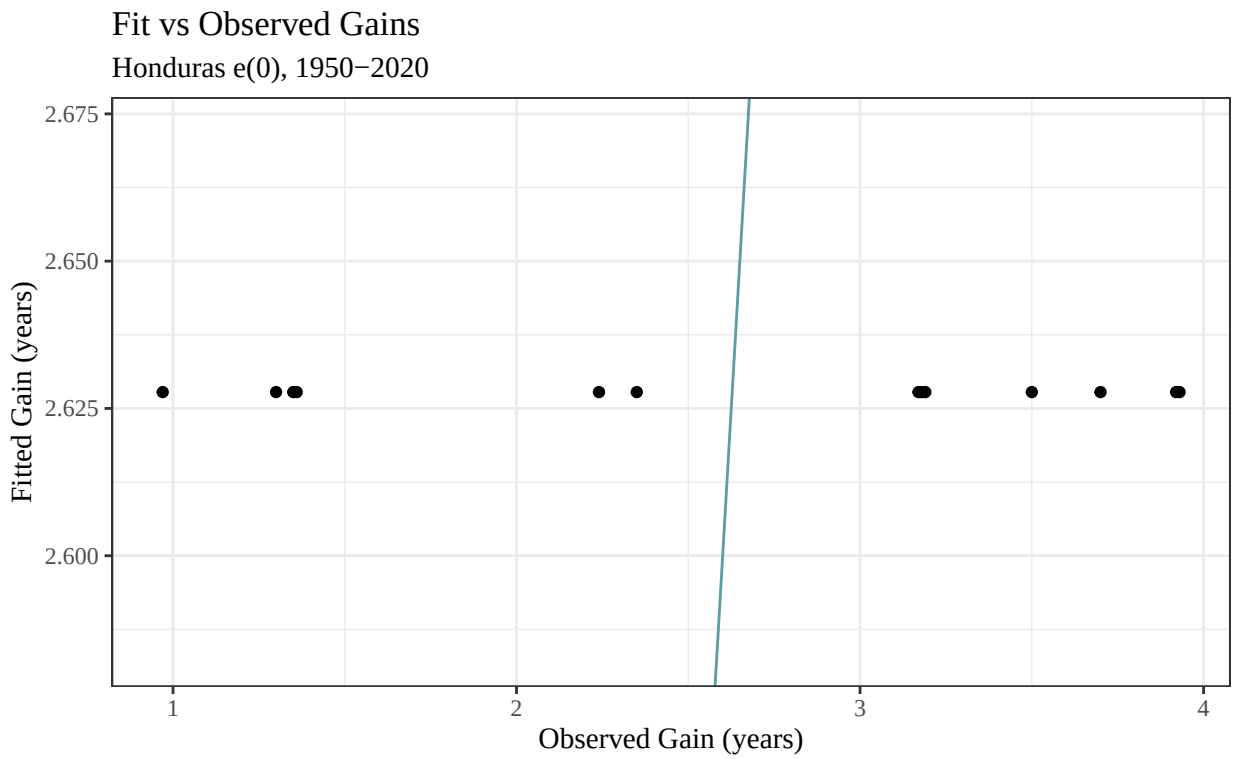


From this plot we choose the starting values: $[3, 3, 4, 2, 1, 4]$, where the fist four items are the deltas, followed by k and z.

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 $[3.059, 3.29, 4.404, 2.293, 1.676, 2.628]$ with an error variance of 14.17.

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:



I'm seeing a weird problem here where my gain function seems to predict the same value no matter the input e_0, not sure what's going on here.

Q1.c

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 from our model:

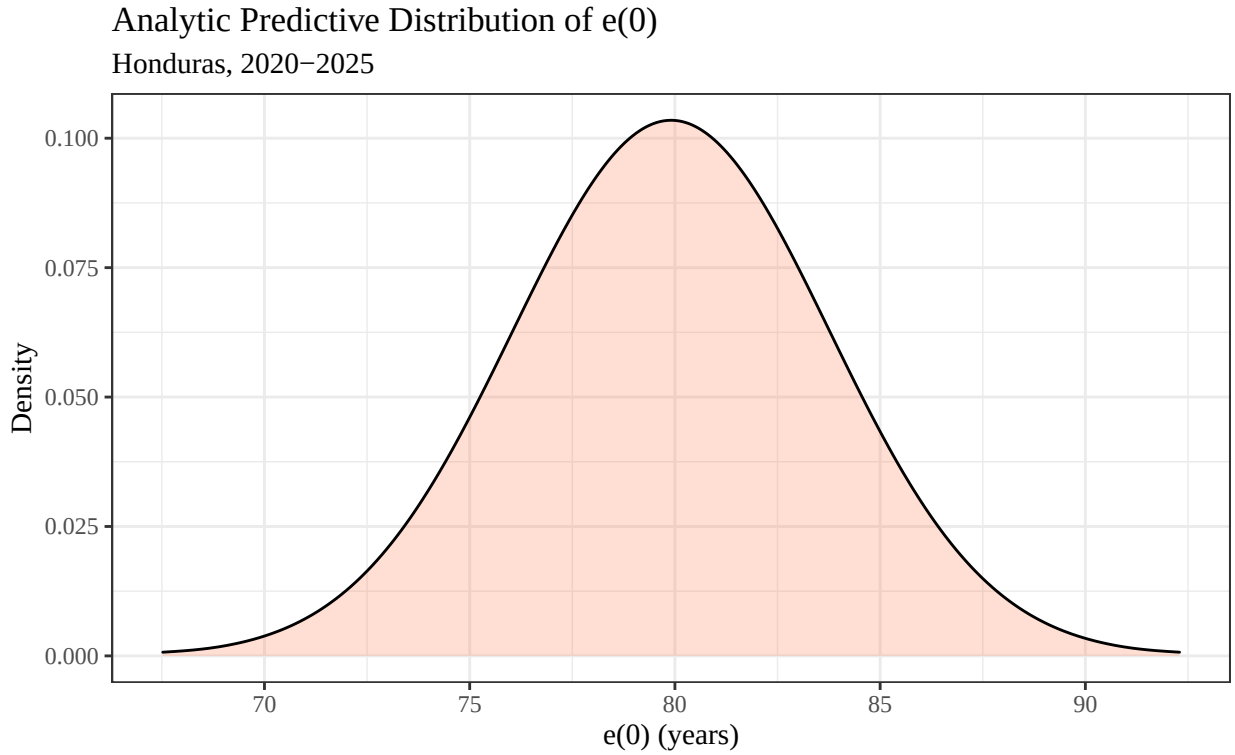


Table 2: Summary of the predictive distribution of 2020-2025 Honduras e_0

Mean	Median	2.5% PI	97.5% PI
79.908	79.908	72.53	87.286

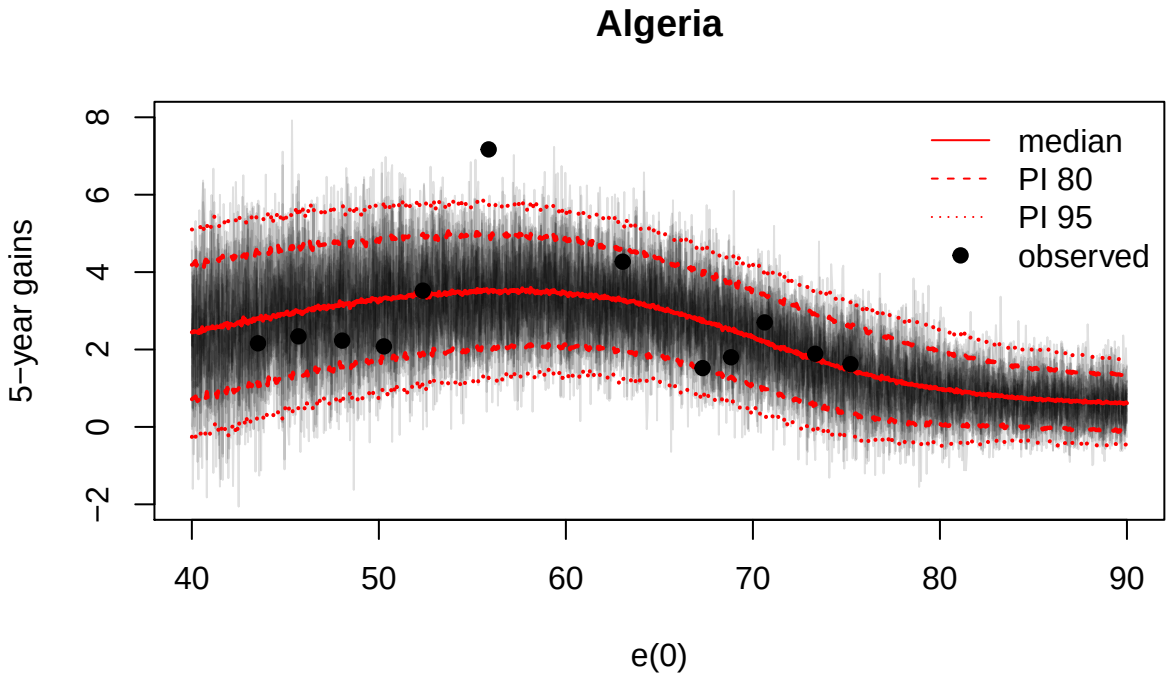
Q2

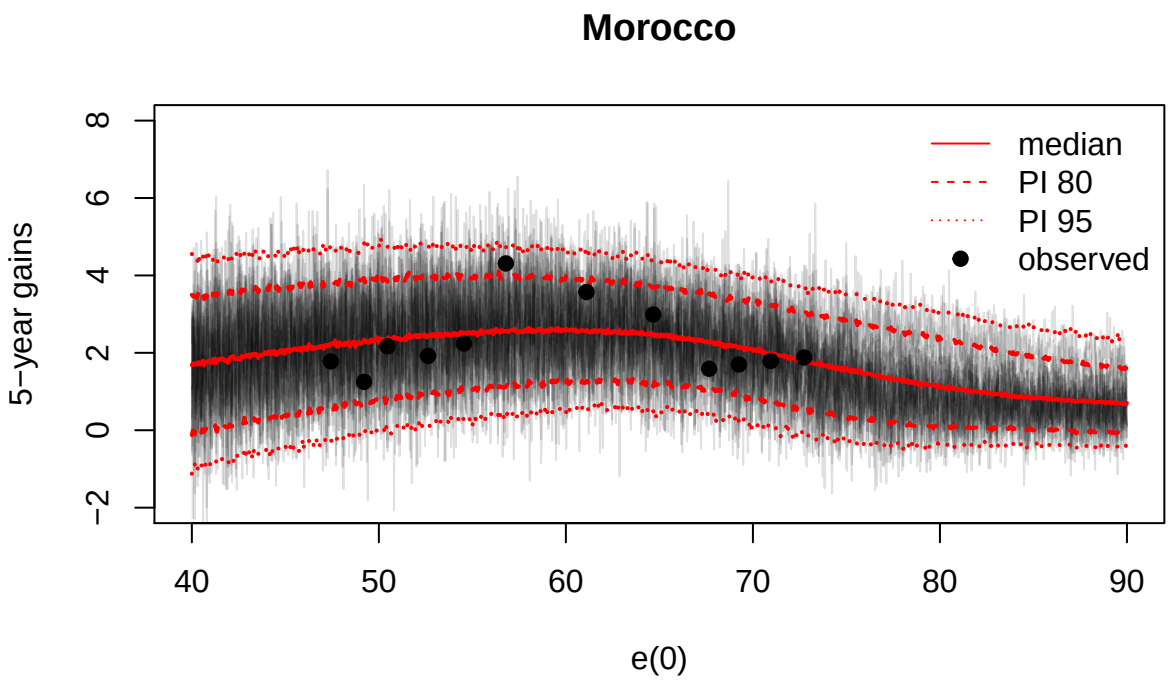
Q2.a

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 projections created using an input set of MCMC chains.

Q2.b

Double logistic curve fits for Algeria and Morocco:





From these double logistic curve fits, we see that both countries reach their peak 5-year gains around when life expectancy at birth is 60 years, though Algeria has a higher expected 5-year gains on average than Morocco for a given e_0. Morocco also has a wider probability interval for e_0 above 70.

Q2.c

Looking at the double-logistic plots, it appears that Algeria has been experiencing faster increases in life expectancy at birth, as reported by the larger average predicted gains at all values of e_0. We can confirm this by examining the reported life expectancies for each country from 1950 to 2015:

Table 3: e_0 for Algeria and Morocco, 1950-2015

Period start	Algeria	Morocco
1950	43.54	47.43
1955	45.70	49.21
1960	48.04	50.46
1965	50.27	52.63
1970	52.35	54.55
1975	55.87	56.79
1980	63.04	61.10
1985	67.31	64.67
1990	68.83	67.66
1995	70.63	69.25
2000	73.33	70.95
2005	75.22	72.74
2010	76.84	74.62

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

Table 4: Mean gains in e_0 over 1950-2015 for Algeria and Morocco

Country	Mean Gain
Algeria	2.775
Morocco	2.266

which confirms the previous two observations.

Q2.d

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

Table 5: Probability of Algeria having a higher average e_0 in each future period

Period start	Pr(DZA > MAR)	PR(DZA < MAR)
2015	0.930	0.070
2020	0.803	0.197
2025	0.699	0.301
2030	0.635	0.365
2035	0.597	0.403
2040	0.572	0.428

Period start	Pr(DZA > MAR)	PR(DZA < MAR)
2045	0.532	0.468
2050	0.526	0.474
2055	0.507	0.493
2060	0.502	0.498
2065	0.492	0.508
2070	0.492	0.508
2075	0.475	0.525
2080	0.466	0.534
2085	0.465	0.535
2090	0.462	0.538
2095	0.460	0.540

Similarly, we can find the probability of e_0 either being higher or lower in Algeria than Morocco in all future periods by finding the fraction of simulations where each condition is met:

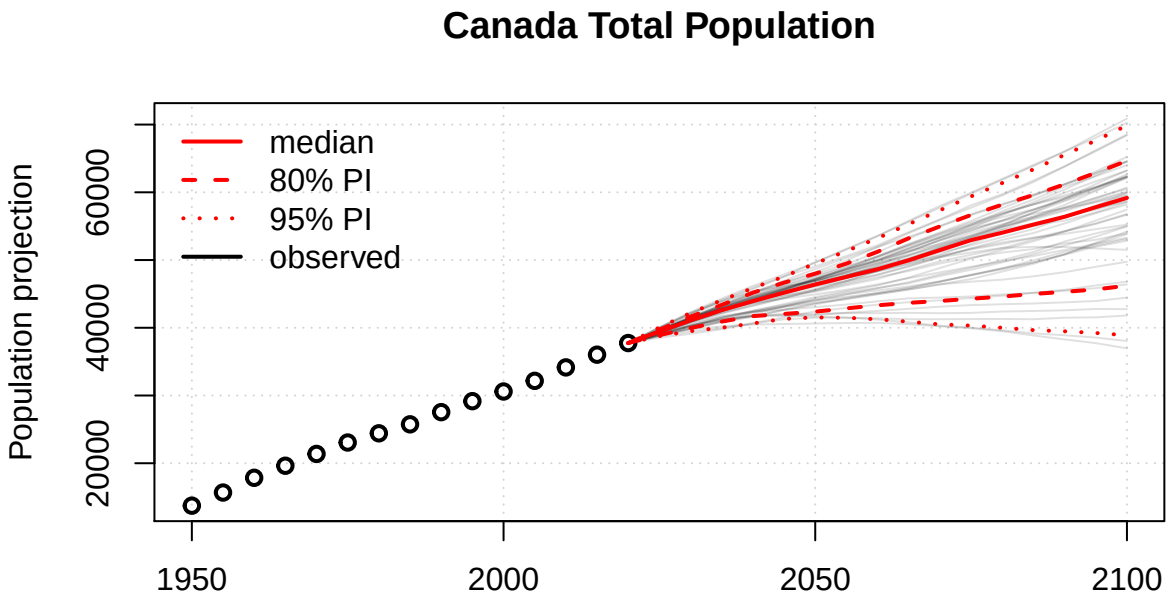
$$Pr(\text{DZA} > \text{MAR}) = 0.347$$
$$Pr(\text{DZA} < \text{MAR}) = 0.049$$

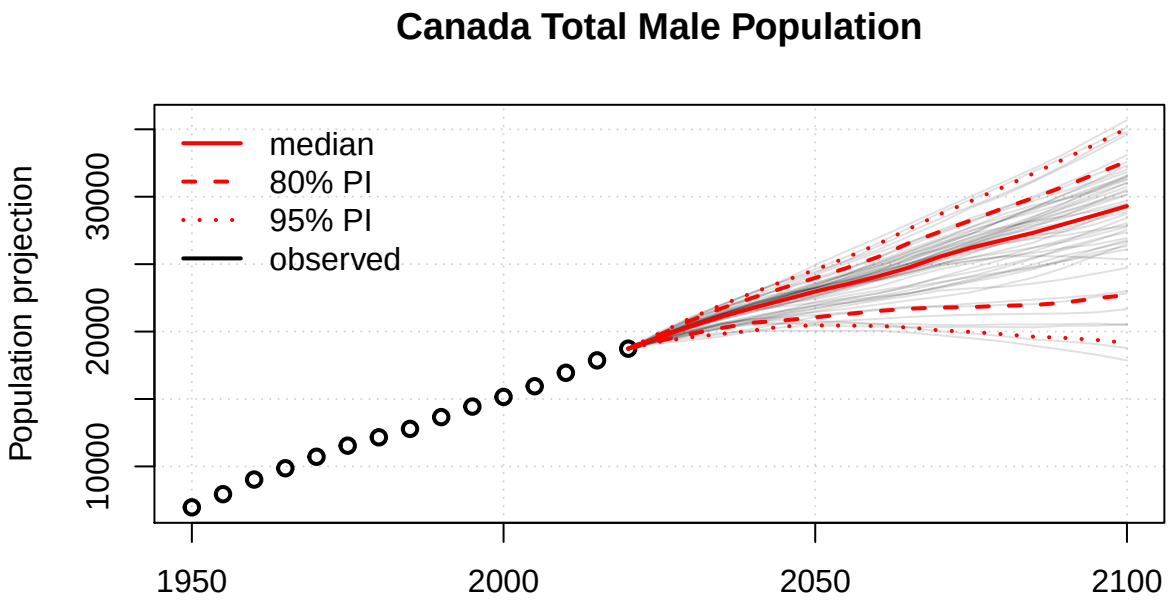
Q3

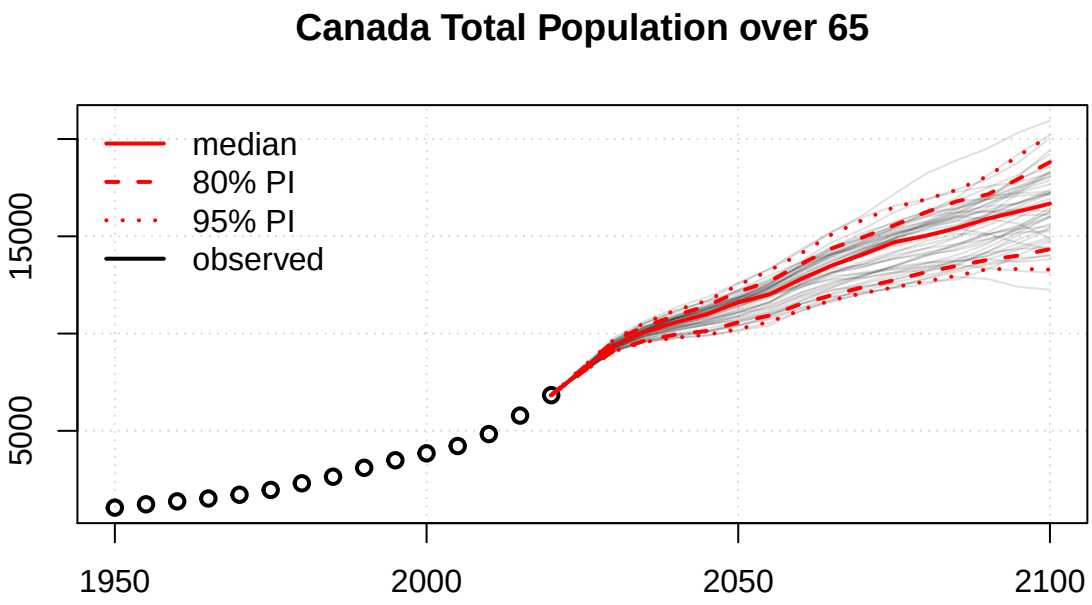
Q3.a

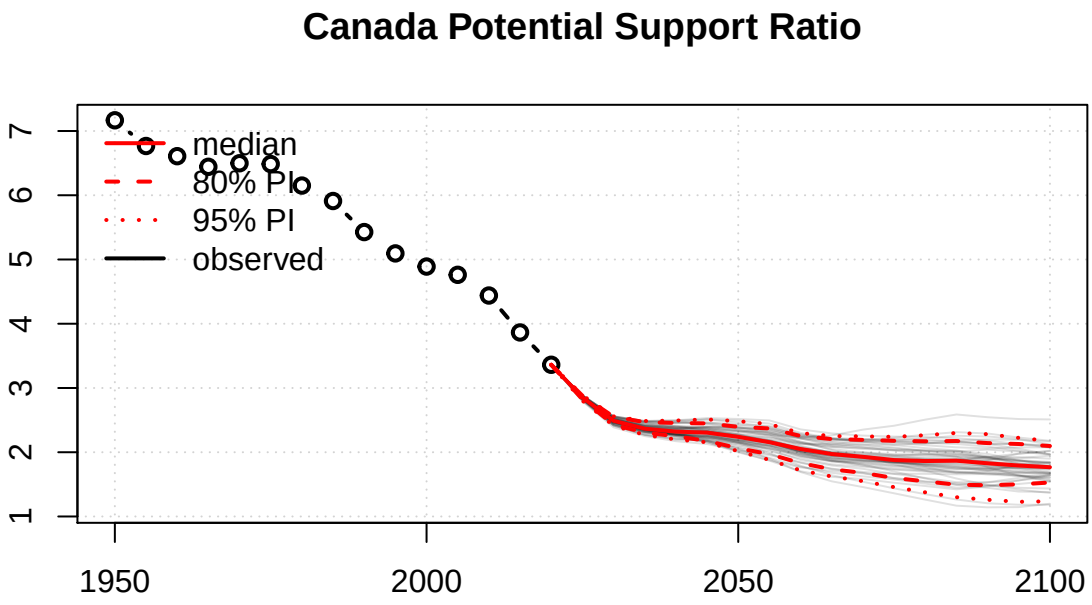
Using our converged life expectancy at birth and total fertility rate (from homework 5) simulations, we can create probabilistic projections of the following population quantities:

Note that the "Potential Support Ratio" is defined as $\frac{\textit{people aged 20-64}}{\textit{people aged 65 and over}}$



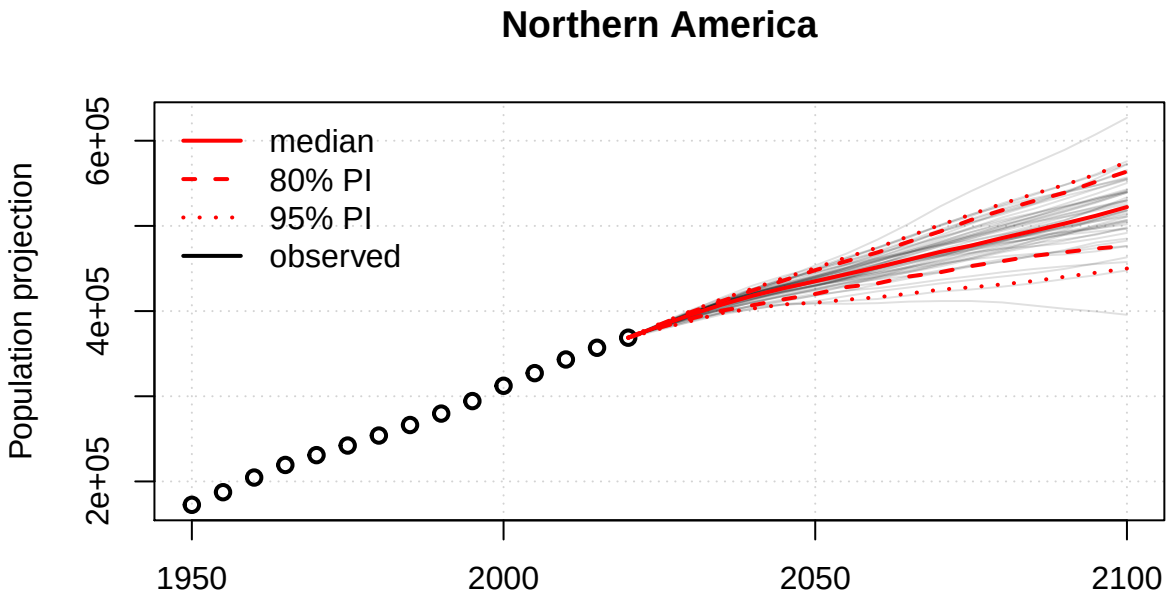

Canada Total Male Population
Population projection
10000
20000
30000
median
80% PI
95% PI
observed
1950
2000
2050
2100


Canada Total Population over 65
5000
15000
median
80% PI
95% PI
observed
1950
2000
2050
2100



Looking at the projections of potential support ratio, we see continued declines into the future, though eventually at a slower rate. This indicates Canada's age structure is likely to continue shifting to older ages.

Q3.b



We can also aggregate the projected population simulations from multiple countries to get the probabilistic population projection of a region, like North America.

Q4

Q4.a

Table 6: Crude net migration rate (CNMR) for Canada, 1950-2100

Period start	CNMR
1950	7.560
1955	6.740
1960	1.907
1965	5.962
1970	6.825
1975	3.544
1980	2.982
1985	6.450
1990	4.871
1995	5.245
2000	6.706
2005	8.001
2010	7.076
2015	6.562

Q4.b

After fitting an autoregressive ($AR(1)$) model to the series of crude net migration rates over time, we get a model with the parameters:

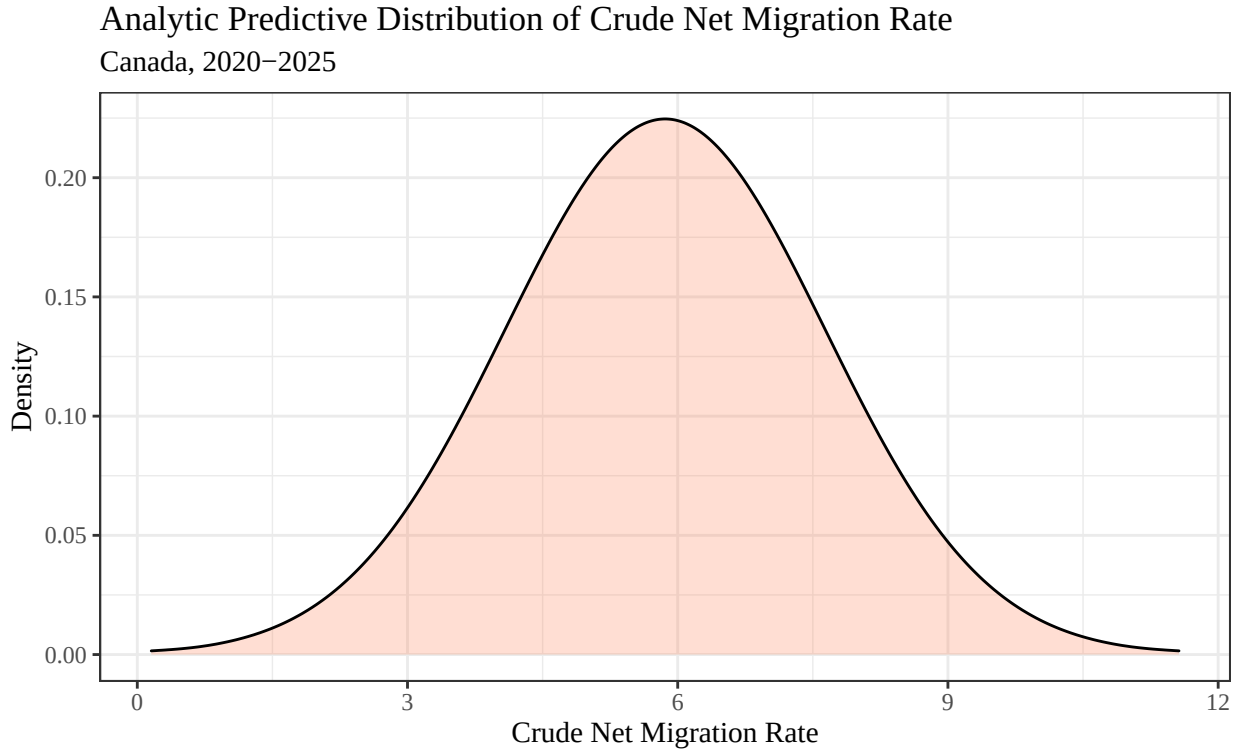
Table 7: AR(1) model parameters for Canada CNMR, 1950-2020

Parameter	Value	S.E
AR(1) Param	0.115	0.270
Mean	5.769	0.522

Q4.c

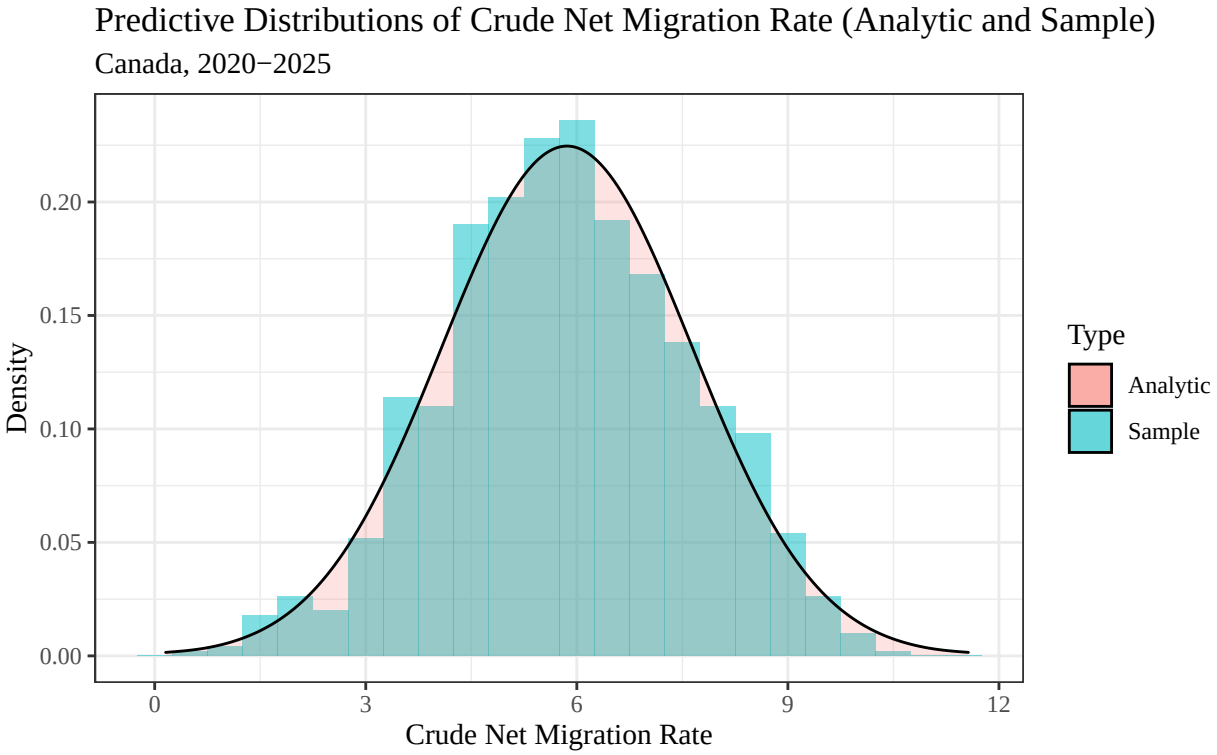
Using these model parameters, we can find an analytic solution for the predictive probability distribution of net migration rates in Canada for 2020-2025, which takes the form:

$$\text{CNMR}_{2020-2025} \sim N(5.86, 1.734)$$



Q4.d

We can also sample from this distribution to show it does indeed follow what we predict analytically:



Q4.e

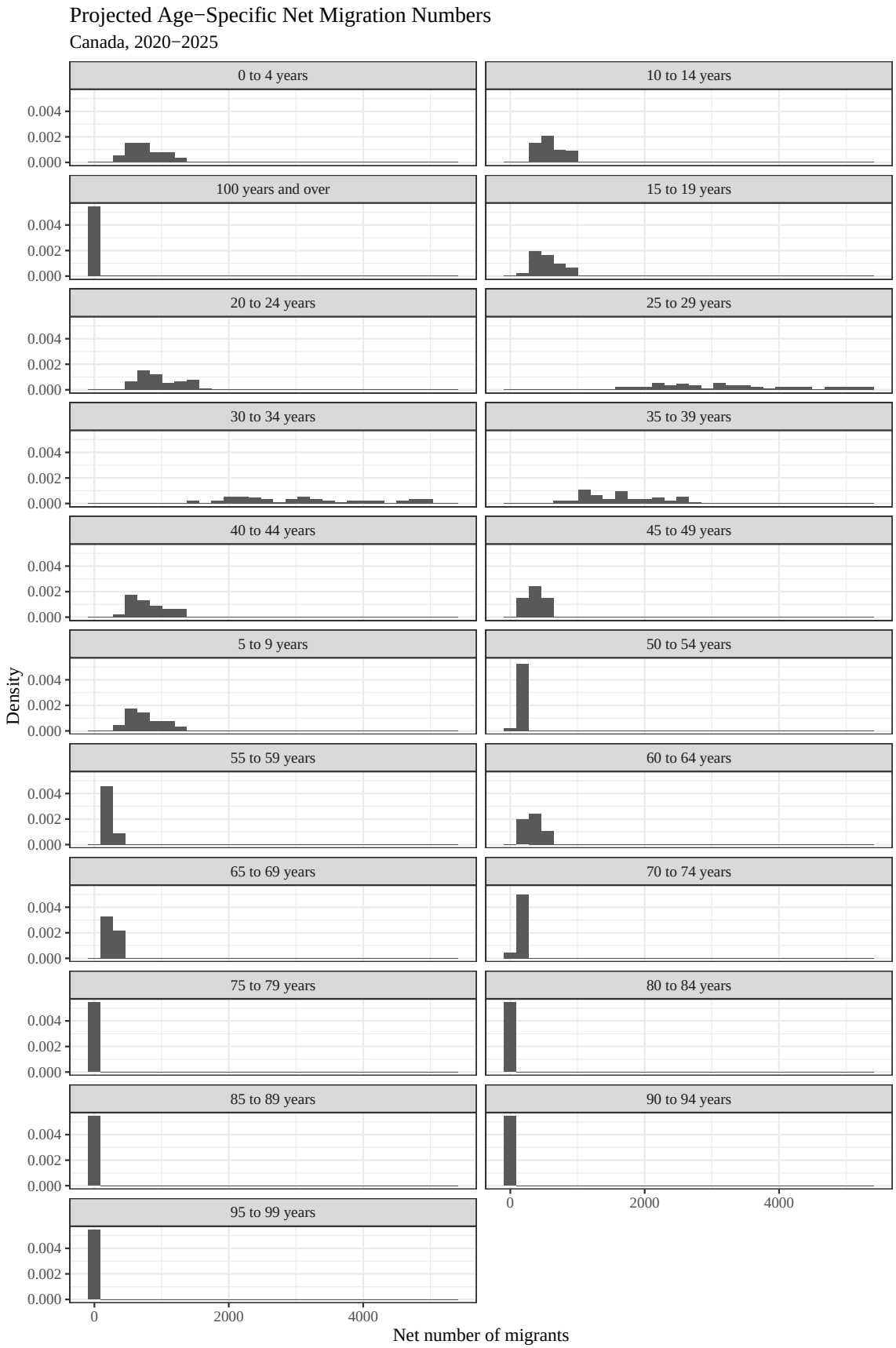
Using data downloaded from Canada's statistical website (*www.statcan.gc.ca*), we can get the age-specific net migration number for 2018-2019. From this, we can then get the proportion of net migration in each age group. *Note that there are zero values because some age groups had more emigrants than immigrants. These were coerced to zero, under the assumption that old ages are less likely to migrate.*

Table 8: Age schedule of net migration rate (per 1000 people)

Age group	Net migration rate
0 to 4 years	0.394
5 to 9 years	0.389
10 to 14 years	0.306
15 to 19 years	0.273
20 to 24 years	0.452
25 to 29 years	1.319
30 to 34 years	1.115
35 to 39 years	0.629
40 to 44 years	0.306
45 to 49 years	0.157
50 to 54 years	0.072
55 to 59 years	0.086
60 to 64 years	0.131
65 to 69 years	0.112
70 to 74 years	0.075
75 to 79 years	0.035
80 to 84 years	0.011
85 to 89 years	0.000

Age group	Net migration rate
90 to 94 years	0.000
95 to 99 years	0.000
100 years and over	0.000

Q4.f


Projected Age−Specific Net Migration Numbers
Canada, 2020−2025
0 to 4 years
10 to 14 years
100 years and over
15 to 19 years
20 to 24 years
25 to 29 years
30 to 34 years
35 to 39 years
40 to 44 years
45 to 49 years
5 to 9 years
50 to 54 years
55 to 59 years
60 to 64 years
65 to 69 years
70 to 74 years
75 to 79 years
80 to 84 years
85 to 89 years
90 to 94 years
95 to 99 years
Density
0.004
0.002
0.000
0
2000
4000
Net number of migrants

Using our 50 simulated population projections, we can get the probabilistic projected age-specific net migration numbers by taking a samples from the $AR(1)$ model and applying the same schedule of age-specific migration for each simulated trajectory.

Q4.g

Appendix

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

# Load libraries
library(dplyr)
library(tidyr)
library(tibble)
library(ggplot2)
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(9876)


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

e0f_hnd <- e0f_all %>%
  filter(name == "Honduras") %>%
  select(-name) %>%
  mutate(gain = lead(e0) - e0)

knitr::kable(
  e0f_hnd,
  booktabs = TRUE, digits = 3, eval = FALSE,
  col.names = c("Period start", "$e_0$", "Gain"),
```

```
  caption = "Honduras female life expectancy at birth and observed gains, 1950-2020"
)


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

dl_gain <- function(l, theta) {

  d1 <- theta[1]
  d2 <- theta[2]
  d3 <- theta[3]
  d4 <- theta[4]

  k <- theta[5]
  z <- theta[6]

  (
    (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) )))
  )

}

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)
e0.DLcurve.plot(e0_sim_mcmc, country = "Honduras")

opt_input <- e0f_hnd %>% slice(-n())
loss_func <- ls_err(dl_gain, opt_input$e0, opt_input$gain)

starting_params <- c(3, 3, 4, 2, 1, 4)
opt_result <- optim(starting_params, loss_func)

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


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

e0f_hnd_gains <- e0f_hnd %>%
  rename(obs_gain = gain) %>%
  mutate(fit_gain = dl_gain(e0, opt_params))

ggplot(e0f_hnd_gains, aes(x = obs_gain, y = fit_gain)) +
```

```r
  geom_point() +
  geom_abline(slope = 1, intercept = 0, color = "cadetblue") +
  theme_bw() +
  theme(text = element_text(family = "serif")) +
  labs(
    title = "Fit vs Observed Gains",
    subtitle = "Honduras e(0), 1950-2020",
    x = "Observed Gain (years)",
    y = "Fitted Gain (years)"
  )


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

hnd_e0_2020_mean <- e0f_hnd_gains %>%
  filter(year == 2015) %>%
  select(e0, fit_gain) %>%
  rowSums()

hnd_e0_2020_sd <- sqrt(opt_err_var)

hnd_e0_2020_dist <- qnorm(
  seq(.0005, .9995, .001),
  mean = hnd_e0_2020_mean,
  sd = hnd_e0_2020_sd
)

ggplot(enframe(hnd_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 = "Honduras, 2020-2025",
    x = "e(0) (years)",
    y = "Density"
  )

hnd_e0_2020_tbl <- tibble(
  Mean = hnd_e0_2020_mean,
  Median = median(hnd_e0_2020_dist),
  `2.5% PI` = Mean - 1.96 * hnd_e0_2020_sd,
  `97.5% PI` = Mean + 1.96 * hnd_e0_2020_sd
)

knitr::kable(
  hnd_e0_2020_tbl,
  booktabs = TRUE, digits = 3, eval = FALSE,
  caption = "Summary of the predictive distribution of 2020-2025 Honduras $e_0$"
)


# Question 2 -------------------------------------------------------------
```

```r
# Question 2a ------------------------------------------------------------

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


# Question 2b ------------------------------------------------------------

e0.DLcurve.plot(e0_sim_mcmc, "Algeria",
  predictive.distr = TRUE, pi = c(80, 95), ylim = c(-2, 8)
)

e0.DLcurve.plot(e0_sim_mcmc, "Morocco",
  predictive.distr = TRUE, pi = c(80, 95), ylim = c(-2, 8)
)


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

e0_traj_tbl <- list(Algeria = "Algeria", Morocco = "Morocco") %>%
  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)) %>%
  arrange(desc(`Mean Gain`))

knitr::kable(
  e0_traj_tbl,
  booktabs = TRUE, digits = 3, eval = FALSE,
  caption = "$e_0$ for Algeria and Morocco, 1950-2015"
)
knitr::kable(
  e0_traj_mean_tbl,
  booktabs = TRUE, digits = 3, eval = FALSE,
  caption = "Mean gains in $e_0$ over 1950-2015 for Algeria and Morocco"
)

# Question 2d ------------------------------------------------------------

e0_dza_traj <- get.e0.trajectories(e0_sim_pred, "Algeria")[-1, ]
e0_mar_traj <- get.e0.trajectories(e0_sim_pred, "Morocco")[-1, ]

dza_gt_mar_each <- rowMeans(e0_dza_traj > e0_mar_traj)
dza_lt_mar_each <- rowMeans(e0_dza_traj < e0_mar_traj)
```

```r
dza_gt_mar_all <- mean(apply(e0_dza_traj > e0_mar_traj, 2, all))
dza_lt_mar_all <- mean(apply(e0_dza_traj < e0_mar_traj, 2, all))

dza_mar_prob_each_tbl <- tibble(
  period_start = as.integer(names(dza_gt_mar_each)) - 3,
  prob_dza_higher = dza_gt_mar_each,
  prob_dza_lower = dza_lt_mar_each
)

knitr::kable(
  dza_mar_prob_each_tbl,
  booktabs = TRUE, digits = 3, eval = FALSE,
  col.names = c("Period start", "Pr(DZA > MAR)", "PR(DZA < MAR)"),
  caption = "Probability of Algeria having a higher average $e_0$ in each future period"
)

# Question 3 ------------------------------------------------------------

# Run this once to get pop predictions

# pop_sim_pred <- pop.predict(
#   end.year = 2100, start.year = 1950, present.year = 2020, wpp.year = 2019,
#   output.dir = pop_sim_dir,
#   inputs = list(
#     tfr.sim.dir = tfr_sim_dir,
#     e0F.sim.dir = e0_sim_dir,
#     e0M.sim.dir = "joint_"
#   ),
#   nr.traj = 50,
#   keep.vital.events = FALSE
# )

pop_sim_pred <- get.pop.prediction(pop_sim_dir)


# Question 3a -----------------------------------------------------------

# Defined using `?pop.expressions`
over65_exp <- "PCAN[14:27]"
support_exp <- "PCAN[5:13] / PCAN[14:27]"

pop.trajectories.plot(
  pop_sim_pred, "Canada",
  sex = "both",
  sum.over.ages = TRUE,
  main = "Canada Total Population"
)

pop.trajectories.plot(
  pop_sim_pred, "Canada",
  sex = "male",
  sum.over.ages = TRUE,
  main = "Canada Total Male Population"
```

```
)

pop.trajectories.plot(
  pop_sim_pred, "Canada",
  expression = over65_exp,
  sex = "both",
  sum.over.ages = TRUE,
  main = "Canada Total Population over 65"
)

pop.trajectories.plot(
  pop_sim_pred, "Canada",
  expression = support_exp,
  sex = "both",
  sum.over.ages = TRUE,
  main = "Canada Potential Support Ratio"
)

# Question 3b ------------------------------------------------------------

# pop_agg_n_am <- pop.aggregate(pop_sim_pred, regions = 905)
pop_agg_n_am <- get.pop.aggregation(pop_sim_dir)

pop.trajectories.plot(pop_agg_n_am, 905, sum.over.ages = TRUE)


# Question 4 -------------------------------------------------------------

can_mig <- readr::read_csv(mig_file) %>%
  filter(Location == "Canada") %>%
  transmute(year = MidPeriod - 3, cnmr = CNMR) %>%
  filter(year < 2020)


# Question 4a ------------------------------------------------------------

knitr::kable(
  can_mig,
  booktabs = TRUE, digits = 3,
  col.names = c("Period start", "CNMR"),
  caption = "Crude net migration rate (CNMR) for Canada, 1950-2100"
)


# Question 4b ------------------------------------------------------------

can_mig_ar_model <- arima(can_mig$cnmr, order = c(1, 0, 0), method = "ML")

can_mig_ar_tbl <- tibble(
  Parameter = c("AR(1) Param", "Mean"),
  Value = can_mig_ar_model$coef,
  `S.E` = sqrt(diag(vcov(can_mig_ar_model)))
)
```

```r
knitr::kable(
  can_mig_ar_tbl,
  booktabs = TRUE, digits = 3,
  caption = "AR(1) model parameters for Canada CNMR, 1950-2020"
)

# Question 4c ------------------------------------------------------------

can_r_2015 <- can_mig %>% filter(year == 2015) %>% pull(cnmr)
can_mu <- can_mig_ar_model$coef[[2]]
can_phi <- can_mig_ar_model$coef[[1]]

can_r_2020_mean <- can_phi * (can_r_2015 - can_mu) + can_mu
can_r_2020_sd <- sqrt(can_mig_ar_model$sigma2)

can_r_2020_dist <- qnorm(
  seq(.0005, .9995, .001),
  mean = can_r_2020_mean,
  sd = can_r_2020_sd
)

ggplot(enframe(can_r_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 Crude Net Migration Rate",
    subtitle = "Canada, 2020-2025",
    x = "Crude Net Migration Rate",
    y = "Density"
  )


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

can_r_2020_sample <- rnorm(1000, mean = can_r_2020_mean, sd = can_r_2020_sd)

can_r_2020_tbl <-
  tibble(
    Analytic = can_r_2020_dist,
    Sample = can_r_2020_sample
  ) %>%
  pivot_longer(everything(), names_to = "dist", values_to = "value")

ggplot(can_r_2020_tbl, aes(x = value, fill = dist)) +
  geom_histogram(
    data = filter(can_r_2020_tbl, dist == "Sample"),
    aes(y = ..density..),
    binwidth = .5,
    alpha = .5
  ) +
  geom_density(
    data = filter(can_r_2020_tbl, dist == "Analytic"),
```

```
    alpha = .2
  ) +
  theme_bw() +
  theme(text = element_text(family = "serif")) +
  labs(
    title = "Predictive Distributions of Crude Net Migration Rate (Analytic and Sample)",
    subtitle = "Canada, 2020-2025",
    x = "Crude Net Migration Rate",
    y = "Density",
    fill = "Type"
  )


# Question 4e ------------------------------------------------------------

can_mig_num_full <- readr::read_csv(can_mig_num_file) %>%
  select(age_group = `Age group`, sex = `Sex`, type = `Type of migrant`, value = `VALUE`) %>%
  pivot_wider(names_from = "type", values_from = "value") %>%
  mutate(
    net_mig = `Immigrants` - `Emigrants`,
    sex = case_when(
      sex == "Males"      ~ "male",
      sex == "Females"    ~ "female",
      sex == "Both sexes" ~ "both",
      TRUE ~ NA_character_
    )
  ) %>%
  select(-`Immigrants`, -`Emigrants`)

can_mig_prop_both <- can_mig_num_full %>%
  filter(sex == "both") %>%
  select(-sex) %>%
  mutate(
    prop_mig = pmax(net_mig / sum(net_mig), 0),
    mig_rate = can_r_2020_mean * prop_mig
  )

  # extract(age_group, into = "age_group_start", regex = "(\\d+)", convert = TRUE)

knitr::kable(
  select(can_mig_prop_both, age_group, mig_rate),
  booktabs = TRUE, digits = 3,
  col.names = c("Age group", "Net migration rate"),
  caption = "Age schedule of net migration rate (per 1000 people)"
)

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

# Define ages to get from trajectories
age_map_list <- as.list(1:20)
age_map_list[[21]] <- 21:27 # 100+ age aggregate

can_pop_2020_sim <- age_map_list %>%
```

```r
  purrr::map(~pop.trajectories(pop_sim_pred, country = "Canada", age = .x)) %>%
  purrr::map_dfr(~as_tibble(.x, rownames = "year"), .id = "age_id") %>%
  filter(year == 2025) %>%
  mutate(age_group = can_mig_prop_both$age_group) %>%
  select(age_group, everything(), -age_id, -year) %>%
  pivot_longer(-age_group, names_to = "sample", values_to = "pop") %>%
  extract(sample, into = "sample", regex = "(\\d+)", convert = TRUE)

can_mig_prop_both_sample <-
  rnorm(50, mean = can_r_2020_mean, sd = can_r_2020_sd) %>%
  purrr::map_dfr(
    ~mutate(can_mig_prop_both, mig_rate = .x * prop_mig),
    .id = "sample"
  ) %>%
  mutate(sample = as.integer(sample)) %>%
  select(sample, age_group, mig_rate) %>%
  left_join(can_pop_2020_sim, by = c("age_group", "sample")) %>%
  mutate(mig_number = mig_rate * pop)


ggplot(can_mig_prop_both_sample, aes(x = mig_number)) +
  geom_histogram(aes(y = ..density..)) +
  facet_wrap(vars(age_group), ncol = 2) +
  theme_bw() +
  theme(text = element_text(family = "serif")) +
  labs(
    title = "Projected Age-Specific Net Migration Numbers",
    subtitle = "Canada, 2020-2025",
    x = "Net number of migrants",
    y = "Density"
  )

# Question 4g ------------------------------------------------------------

# use pop.predict with migration included?
```