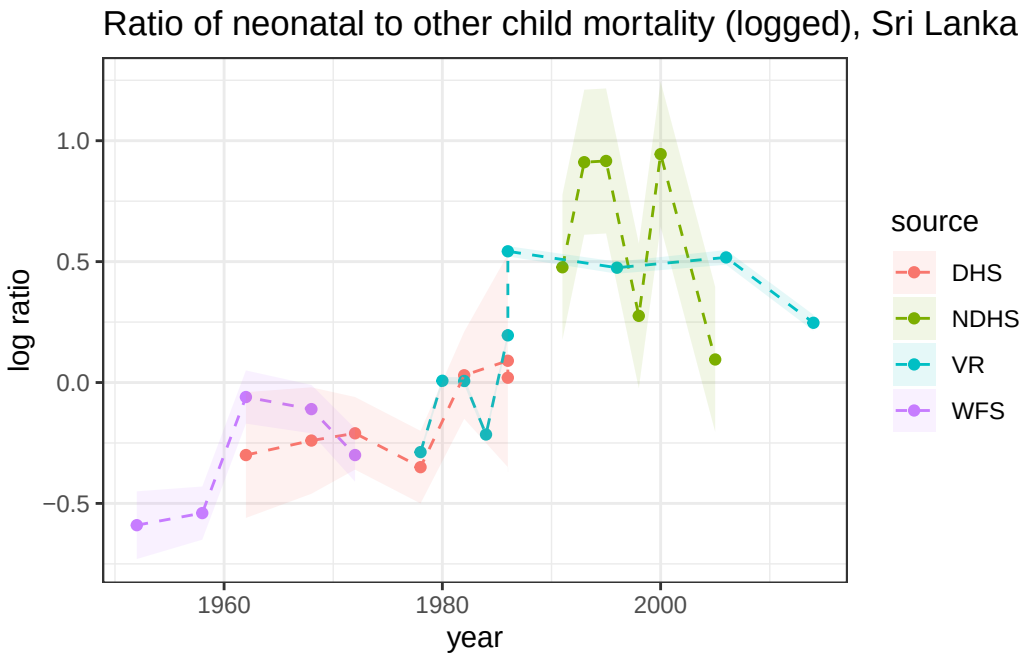
Applied_Stat_Lab_10

24/03/24

Child mortality in Sri Lanka

In this lab you will be fitting a couple of different models to the data about child mortality in Sri Lanka, which was used in the lecture. Here's the data and the plot from the lecture:

```
library(tidyverse)
library(here)
library(rstan)
library(tidybayes)

lka <- read.csv(here("lka.csv"))
ggplot(lka, aes(year, logit_ratio)) +
  geom_point(aes( color = source)) +
  geom_line(aes( color = source), lty = 2) +
  geom_ribbon(aes(ymin = logit_ratio - se,
                  ymax = logit_ratio + se,
                  fill =  source), alpha = 0.1) +
  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka", y = "log
```



Fitting a linear model

Let's firstly fit a linear model in time to these data. Here's the code to do this:

```
observed_years <- lka$year
years <- min(observed_years):max(observed_years)
nyears <- length(years)

stan_data <- list(y = lka$logit_ratio, year_i = observed_years - years[1]+1,
                  T = nyears, years = years, N = length(observed_years),
                  mid_year = mean(years), se = lka$se)

mod <- stan(data = stan_data,
            file = here("code/models/lka_linear_me.stan"))
```

```
SAMPLING FOR MODEL 'anon_model' NOW (CHAIN 1).
Chain 1:
Chain 1: Gradient evaluation took 5.4e-05 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 0.54 seconds.
Chain 1: Adjust your expectations accordingly!
```

```
Chain 1:
Chain 1:
Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 1:
Chain 1:  Elapsed Time: 0.028 seconds (Warm-up)
Chain 1:                0.024 seconds (Sampling)
Chain 1:                0.052 seconds (Total)
Chain 1:

SAMPLING FOR MODEL 'anon_model' NOW (CHAIN 2).
Chain 2:
Chain 2: Gradient evaluation took 3e-06 seconds
Chain 2: 1000 transitions using 10 leapfrog steps per transition would take 0.03 seconds.
Chain 2: Adjust your expectations accordingly!
Chain 2:
Chain 2:
Chain 2: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 2: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 2: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 2: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 2: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 2: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 2: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 2: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 2: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 2: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 2: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 2: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 2:
Chain 2:  Elapsed Time: 0.029 seconds (Warm-up)
Chain 2:                0.027 seconds (Sampling)
Chain 2:                0.056 seconds (Total)
```

```
Chain 2:

SAMPLING FOR MODEL 'anon_model' NOW (CHAIN 3).
Chain 3:
Chain 3: Gradient evaluation took 3e-06 seconds
Chain 3: 1000 transitions using 10 leapfrog steps per transition would take 0.03 seconds.
Chain 3: Adjust your expectations accordingly!
Chain 3:
Chain 3:
Chain 3: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 3: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 3: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 3: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 3: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 3: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 3: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 3: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 3: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 3: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 3: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 3: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 3:
Chain 3:  Elapsed Time: 0.029 seconds (Warm-up)
Chain 3:                0.027 seconds (Sampling)
Chain 3:                0.056 seconds (Total)
Chain 3:

SAMPLING FOR MODEL 'anon_model' NOW (CHAIN 4).
Chain 4:
Chain 4: Gradient evaluation took 3e-06 seconds
Chain 4: 1000 transitions using 10 leapfrog steps per transition would take 0.03 seconds.
Chain 4: Adjust your expectations accordingly!
Chain 4:
Chain 4:
Chain 4: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 4: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 4: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 4: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 4: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 4: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 4: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 4: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 4: Iteration: 1400 / 2000 [ 70%]  (Sampling)
```

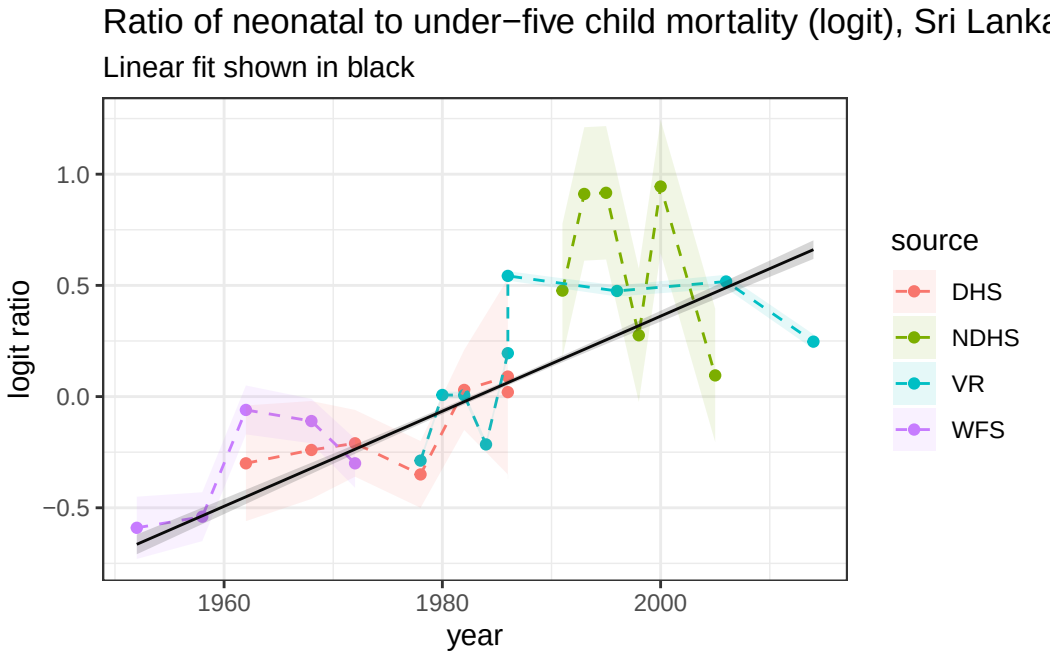
```
Chain 4: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 4: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 4: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 4:
Chain 4:  Elapsed Time: 0.027 seconds (Warm-up)
Chain 4:                0.024 seconds (Sampling)
Chain 4:                0.051 seconds (Total)
Chain 4:
```

Extract the results:

```
res <- mod %>%
  gather_draws(mu[t]) %>%
  median_qi() %>%
  mutate(year = years[t])
```

Plot the results:

```
ggplot(lka, aes(year, logit_ratio)) +
  geom_point(aes( color = source)) +
  geom_line(aes( color = source), lty = 2) +
  geom_ribbon(aes(ymin = logit_ratio - se,
                  ymax = logit_ratio + se,
                  fill =  source), alpha = 0.1) +
  theme_bw()+
  geom_line(data = res, aes(year, .value)) +
  geom_ribbon(data = res, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2)+
  theme_bw()+
  labs(title = "Ratio of neonatal to under-five child mortality (logit), Sri Lanka",
       y = "logit ratio", subtitle = "Linear fit shown in black")
```



Question 1

Project the linear model above out to 2022 by adding a `generated quantities` block in Stan (do the projections based on the expected value μ). Plot the resulting projections on a graph similar to that above.

Answer 1

We fit the model in the following code cell:

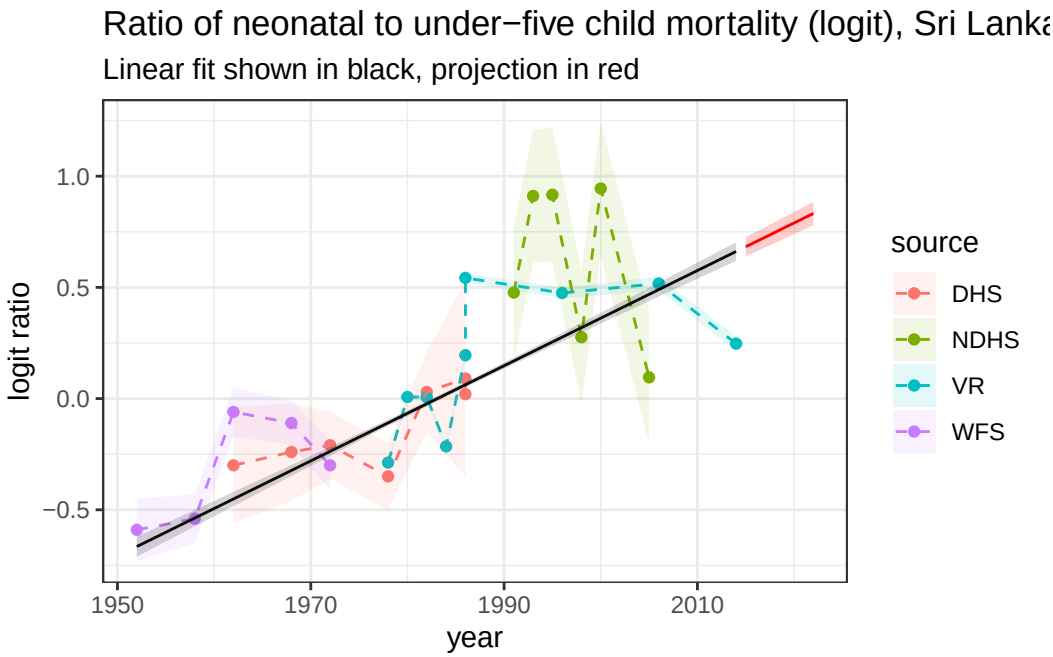
```
stan_data_2 <- list(y = lka$logit_ratio, year_i = observed_years - years[1]+1,
                    T = nyears, years = years, N = length(observed_years),
                    mid_year = mean(years), se = lka$se,P=8)
mod2 <- stan(data = stan_data_2,
             iter = 4000,
             file = here("code/models/lka_linear_me_2.stan"))
```

We then get the draws of the given and projected years:

```
res1 <- mod2 %>%
  gather_draws(mu[t]) %>%
  median_qi() %>%
  mutate(year = years[t])

res1_new <- mod2 %>%
  gather_draws(mu_new[p]) %>%
  median_qi() %>%
  mutate(year = years[nyears]+p)
```

We finally plot the results:

```
ggplot(lka, aes(year, logit_ratio)) +
  geom_point(aes( color = source)) +
  geom_line(aes( color = source), lty = 2) +
  geom_ribbon(aes(ymin = logit_ratio - se,
                  ymax = logit_ratio + se,
                  fill =  source), alpha = 0.1) +
  theme_bw()+
  geom_line(data = res1, aes(year, .value)) +
  geom_ribbon(data = res1, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2)+
  geom_line(data = res1_new, aes(year, .value), col = 'red') +
  geom_ribbon(data = res1_new, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2,
  theme_bw()+
  labs(title = "Ratio of neonatal to under-five child mortality (logit), Sri Lanka",
       y = "logit ratio", subtitle = "Linear fit shown in black, projection in red")
```



Question 2

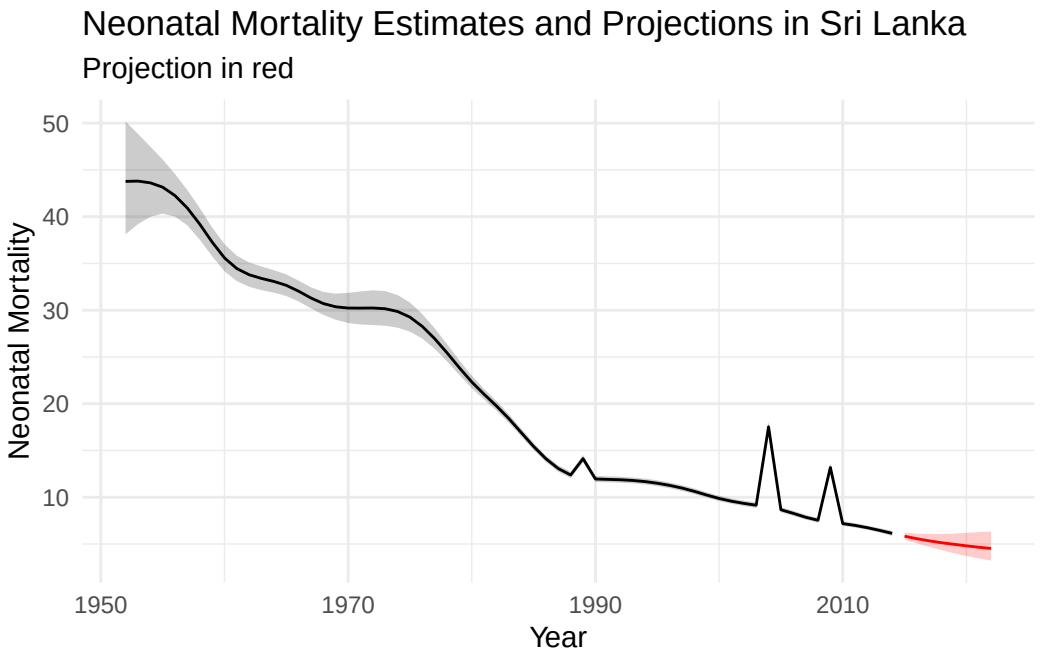
The projections above are for the logit of the ratio of neonatal to under-five child mortality. You can download estimates of the under-five child mortality from 1951 to 2022 here: https://childmortality.org/all-cause-mortality/data/estimates?refArea=LKA. Use these data to get estimates and projections of neonatal mortality for Sri Lanka, and plot the results.

Answer 2

We feed the csv, gather draws and plot the estimates in the following cell:

```
library(janitor)
lka_5 <- read.csv(here("LKAunder5.csv"))
lka_5 <- clean_names(lka_5)

inv_logit <- function(x) {
  exp(x) / (1 + exp(x))
}

ratio_estimate <- rbind(res1 %>% select(.value, .lower, .upper, year),
                        res1_new %>% select(.value, .lower, .upper, year)) %>%
```

```
                        mutate(ratio_est = inv_logit(.value),
                               ratio_lower = inv_logit(.lower),
                               ratio_upper = inv_logit(.upper)
                               )


neo_estimate <- left_join(lka_5, ratio_estimate, by = "year") %>%
                mutate(neo_est = estimate * ratio_est,
                       neo_lower = lower_bound * ratio_lower,
                       neo_upper = upper_bound * ratio_upper)
neo_estimate <- na.omit(neo_estimate)

ggplot(neo_estimate, aes(x = year)) +
  geom_line(data = subset(neo_estimate,year <= 2014),aes(y = neo_est), color = "black") +
  geom_ribbon(data = subset(neo_estimate,year <= 2014) ,aes(ymin = neo_lower, ymax = neo_u
  geom_line(data = subset(neo_estimate,year > 2014),aes(y = neo_est), color = "red") +
  geom_ribbon(data = subset(neo_estimate,year > 2014) ,aes(ymin = neo_lower, ymax = neo_up

  labs(title = "Neonatal Mortality Estimates and Projections in Sri Lanka",
       y = "Neonatal Mortality",
       x = "Year", subtitle = "Projection in red") +
  theme_minimal()
```



Random walks

Question 3

Code up and estimate a first order random walk model to fit to the Sri Lankan data, taking into account measurement error, and project out to 2022.

Answer 3

We fit the model:

```
mod3 <- stan(data = stan_data_2,
             iter = 4000,
             file = here("code/models/lka_linear_me_rw1.stan"))
```

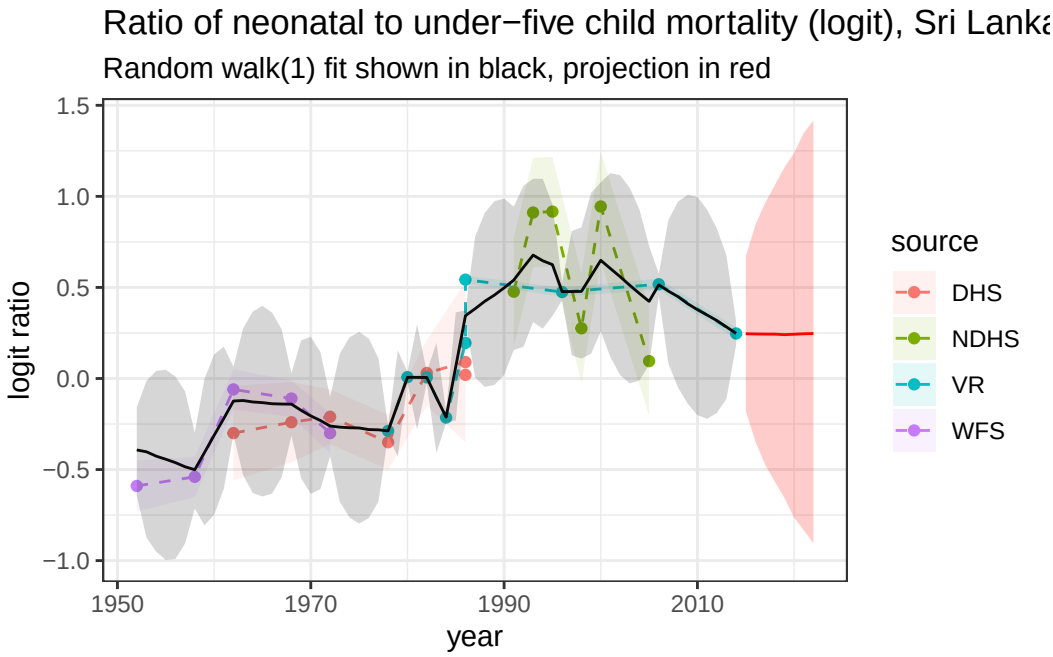
Gather draws:

```
res_rw1 <- mod3 %>%
  gather_draws(mu[t]) %>%
  median_qi() %>%
  mutate(year = years[t])

res_rw1_new <- mod3 %>%
  gather_draws(mu_new[p]) %>%
  median_qi() %>%
  mutate(year = years[nyears]+p)
```

Plot the projection:

```
ggplot(lka, aes(year, logit_ratio)) +
  geom_point(aes( color = source)) +
  geom_line(aes( color = source), lty = 2) +
  geom_ribbon(aes(ymin = logit_ratio - se,
                  ymax = logit_ratio + se,
                  fill =  source), alpha = 0.1) +
  theme_bw()+
  geom_line(data = res_rw1, aes(year, .value)) +
  geom_ribbon(data = res_rw1, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2)+
  geom_line(data = res_rw1_new, aes(year, .value), col = 'red') +
  geom_ribbon(data = res_rw1_new, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0
  theme_bw()+
```

```
  labs(title = "Ratio of neonatal to under-five child mortality (logit), Sri Lanka",
       y = "logit ratio", subtitle = "Random walk(1) fit shown in black, projection in red
```



Question 4

Now alter your model above to estimate and project a second-order random walk model (RW2).

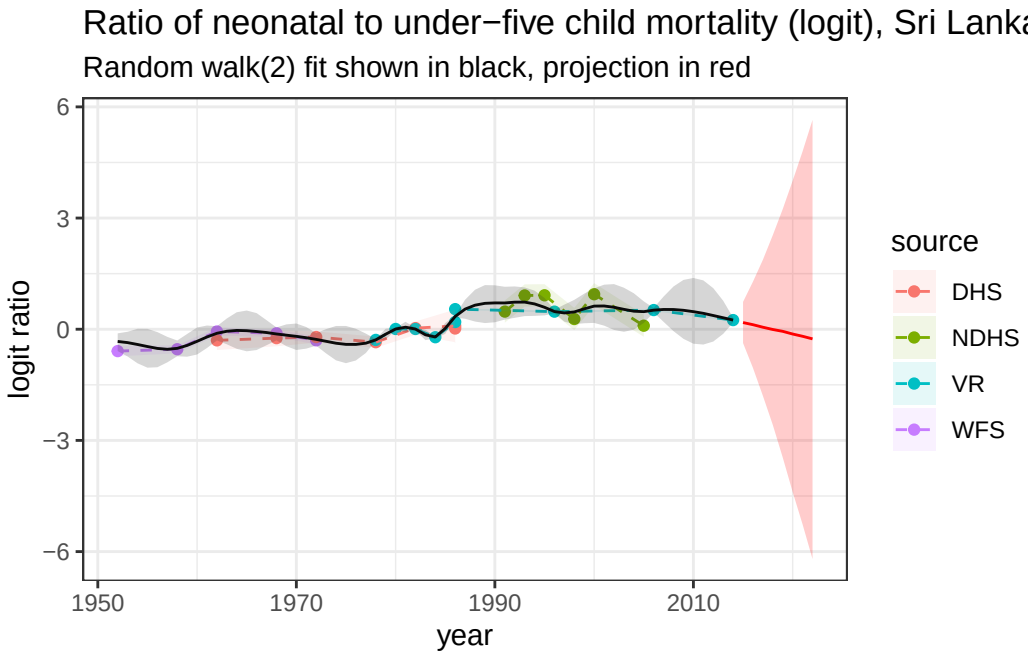
Answer 4

```
mod4 <- stan(data = stan_data_2,
             iter = 4000,
             file = here("code/models/lka_linear_me_rw2.stan"))
```

```
res_rw2 <- mod4 %>%
  gather_draws(mu[t]) %>%
  median_qi() %>%
  mutate(year = years[t])
```

```r
res_rw2_new <- mod4 %>%
  gather_draws(mu_new[p]) %>%
  median_qi() %>%
  mutate(year = years[nyears]+p)
```

```r
ggplot(lka, aes(year, logit_ratio)) +
  geom_point(aes( color = source)) +
  geom_line(aes( color = source), lty = 2) +
  geom_ribbon(aes(ymin = logit_ratio - se,
                  ymax = logit_ratio + se,
                  fill =  source), alpha = 0.1) +
  theme_bw()+
  geom_line(data = res_rw2, aes(year, .value)) +
  geom_ribbon(data = res_rw2, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2)+
  geom_line(data = res_rw2_new, aes(year, .value), col = 'red') +
  geom_ribbon(data = res_rw2_new, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0
  theme_bw()+
  labs(title = "Ratio of neonatal to under-five child mortality (logit), Sri Lanka",
       y = "logit ratio", subtitle = "Random walk(2) fit shown in black, projection in red
```



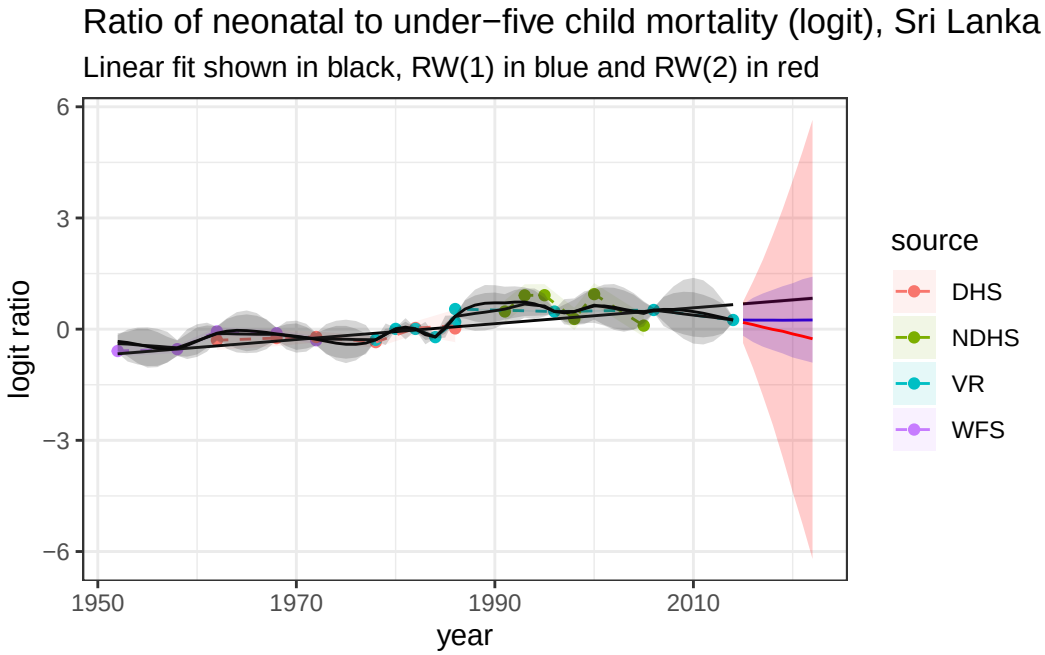
Question 5

Run the first order and second order random walk models, including projections out to 2022. Compare these estimates with the linear fit by plotting everything on the same graph.

Answer 5

We plot all the estimates in the following graph:

```
ggplot(lka, aes(year, logit_ratio)) +
  geom_point(aes( color = source)) +
  geom_line(aes( color = source), lty = 2) +
  geom_ribbon(aes(ymin = logit_ratio - se,
                  ymax = logit_ratio + se,
                  fill =  source), alpha = 0.1) +
  geom_line(data = res1, aes(year, .value)) +
  geom_ribbon(data = res1, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2)+
  geom_line(data = res1_new, aes(year, .value), col = 'black') +
  geom_ribbon(data = res1_new, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2,

  geom_line(data = res_rw1, aes(year, .value)) +
  geom_ribbon(data = res_rw1, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2)+
  geom_line(data = res_rw1_new, aes(year, .value), col = 'blue') +
  geom_ribbon(data = res_rw1_new, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0
  geom_line(data = res_rw2, aes(year, .value)) +
  geom_ribbon(data = res_rw2, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2)+
  geom_line(data = res_rw2_new, aes(year, .value), col = 'red') +
  geom_ribbon(data = res_rw2_new, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0
  theme_bw()+
  labs(title = "Ratio of neonatal to under-five child mortality (logit), Sri Lanka",
       y = "logit ratio", subtitle = "Linear fit shown in black, RW(1) in blue and RW(2) i
```



We have chosen the color 'green' for the linear fit, the color 'blue' for the random walk model of order 1 and the color 'red' for the random walk model of order 2. We see that the logit ratio is projected to increase by the linear model, while the random walk model of order 1 is stationary and random walk model of order 2 shows a decreasing trend, another thing of note is that it has huge lower and upper bounds.

Question 6

Briefly comment on which model you think is most appropriate, or an alternative model that would be more appropriate in this context.

Answer 6

Looking at the plots that we obtained, we see from the estimates over time that it is an overall decreasing trend. However, there is a degree of non-linearity to it presumably due to the quantity being the logit of the ratio and environmental factors. However, we notice the following thing, overall the counts for under 5 year old mortalities and neonatal mortalities are decreasing. So, when we are taking the logit of the ratio, then essentially in the ideal scenario if mortalities go to 0, and the ratio goes to 0 we get the final bound for the logit ratio as :

$$log(\frac{0}{1-0}) = -\infty$$

Therefore, as the second order random walk model captures that decreasing trend in mortality, we select it as the best model for this task.