Missing data and temporal models

Monica Alexander

March 24 2021

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)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(here)
```

```
## here() starts at /Users/siyiwei/Desktop/STA2201_Repo
```

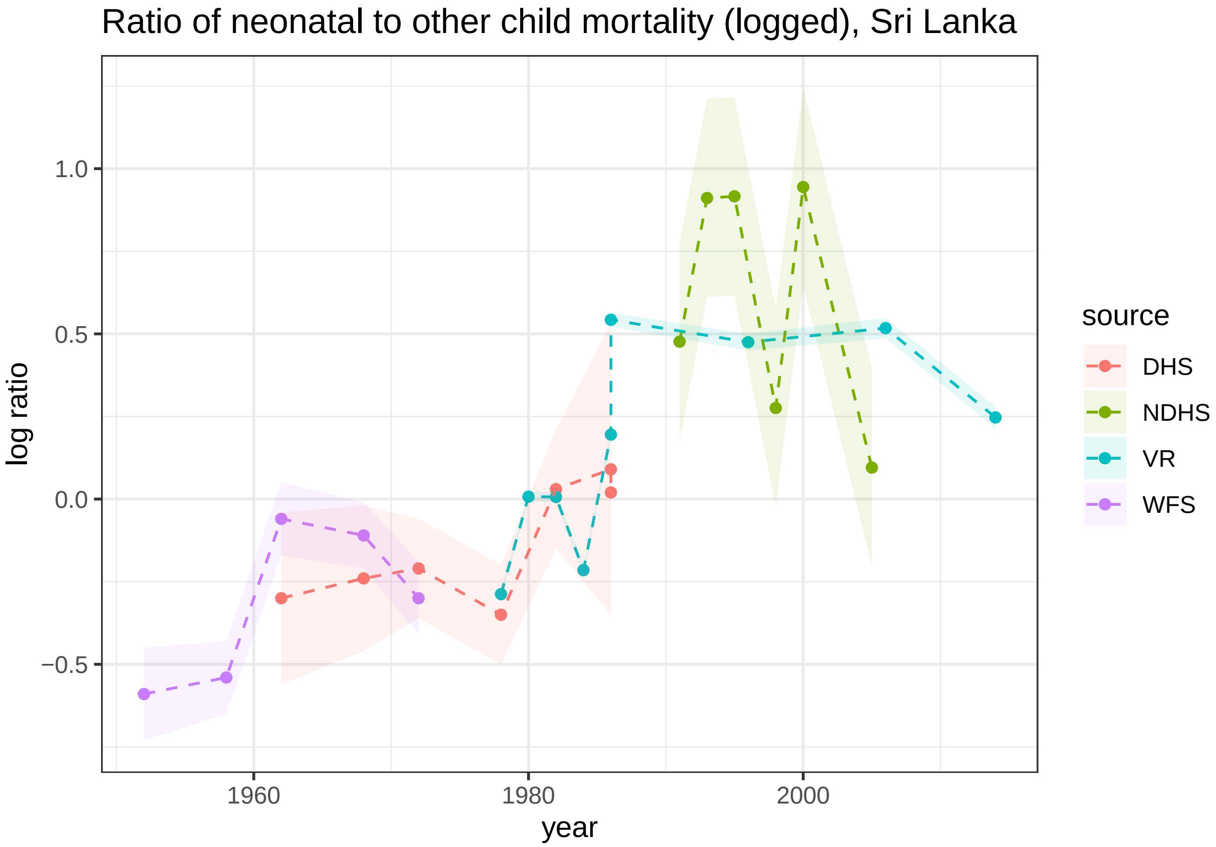
```
library(rstan)
```

```
## Loading required package: StanHeaders

## rstan (Version 2.21.2, GitRev: 2e1f913d3ca3)

## For execution on a local, multicore CPU with excess RAM we recommend calling
## options(mc.cores = parallel::detectCores()).
## To avoid recompilation of unchanged Stan programs, we recommend calling
## rstan_options(auto_write = TRUE)

##
## Attaching package: 'rstan'

## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(tidybayes)

lka <- read_csv(here("data/lka.csv"))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   country_name = col_character(),
##   country_code = col_character(),
##   source = col_character(),
##   year = col_double(),
##   logit_ratio = col_double(),
##   se = col_double(),
##   ratio = col_double()
## )
```

```
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 ratio")
```



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 = "./lka_linear_me.stan")
```

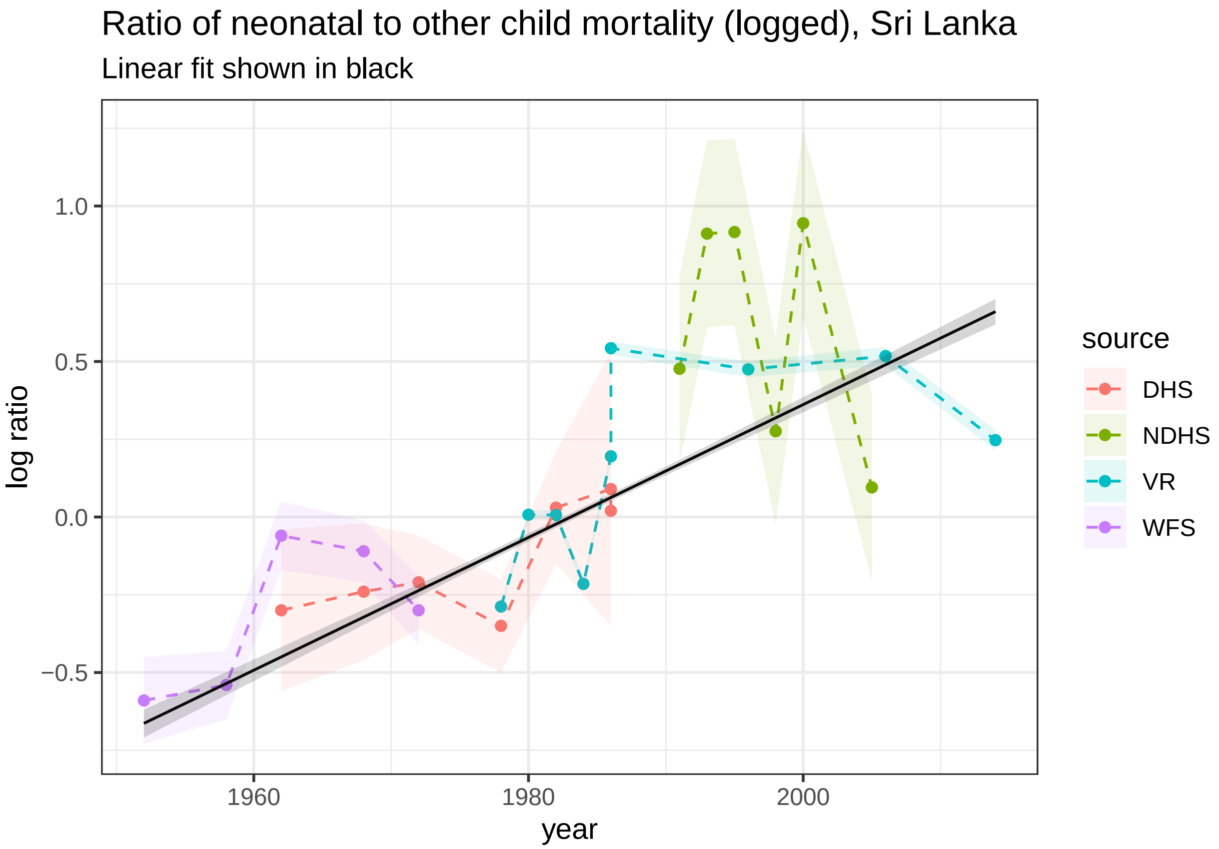
```
## Trying to compile a simple C file
```

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 other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "Linear fit shown in black")
```



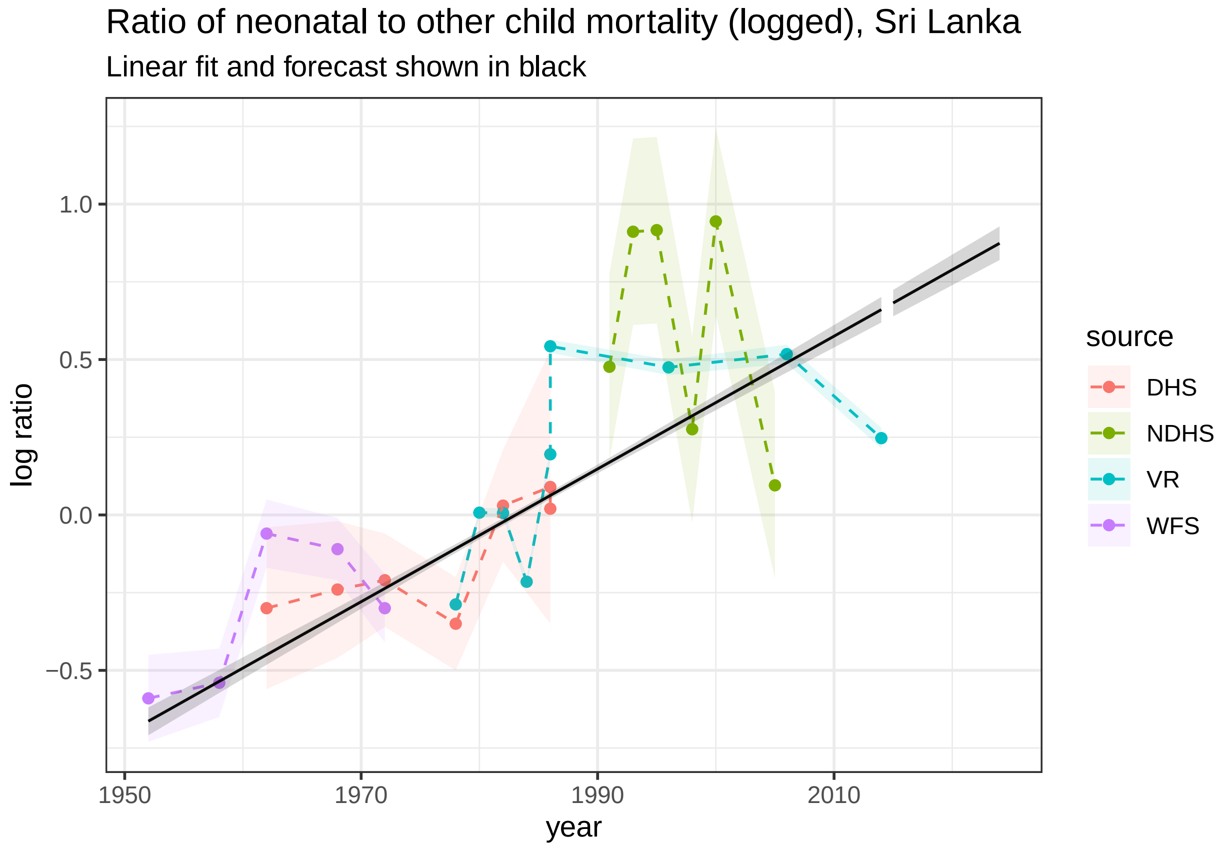
Question 1

Project the linear model above out 10 years past the last observation 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:

The prediction 10 years after could be seen in below graph. Which is still a linear trend.

```
res_pred_linear <- mod %>%
  gather_draws(pred_mu[t]) %>%
  median_qi() %>%
  mutate(year = years[t]+63)

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)+
  geom_line(data = res_pred_linear, aes(year, .value)) +
  geom_ribbon(data = res_pred_linear, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2)+
  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "Linear fit and forecast shown in black")
```



Random walks

The `lka_rw_me` Stan file gives you code to estimate and project a random walk model on the Sri Lankan data.

Question 2

Alter the first order random walk model to estimate and project a second-order random walk model (RW2).

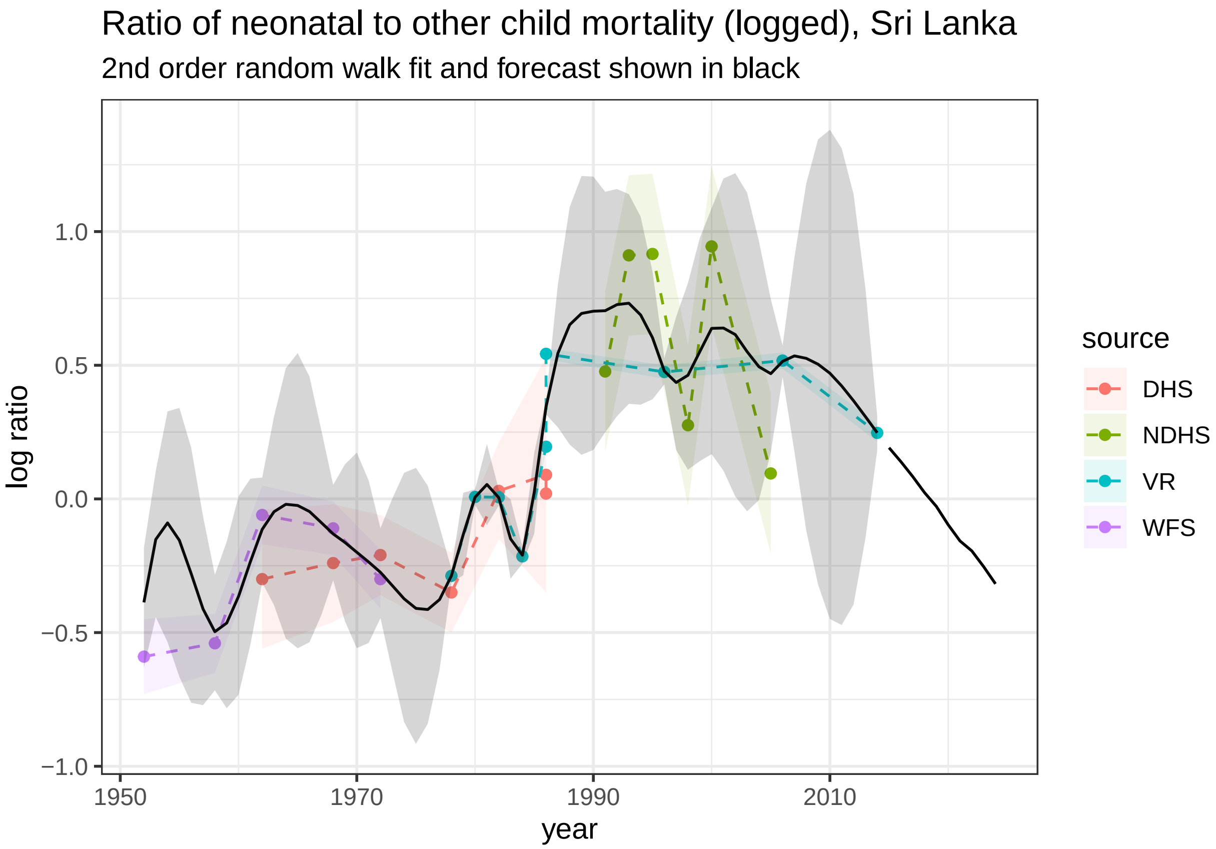
- Answer: I create another model called lka_rw_second which fit the second order random walk. The estimation of second order random walk is shown below. However, like the credible interval shown in the lecture, it is too wide so I decide not to show it.

```
Q3data <- list(y = lka$logit_ratio,
               year_i = observed_years - years[1]+1,
               T = nyears, years = years,
               N = length(observed_years),
               P = 10,
               se = lka$se)
Q3_fit <- stan(file = "./lka_rw_second.stan", data = Q3data)
```

```
## Trying to compile a simple C file
```

```
## Warning: Bulk Effective Samples Size (ESS) is too low, indicating posterior means and medians may be
## Running the chains for more iterations may help. See
## http://mc-stan.org/misc/warnings.html#bulk-ess
```

```
res <- Q3_fit %>%
  gather_draws(mu[t]) %>%
  median_qi() %>%
  mutate(year = years[t])

res_pred <- Q3_fit %>%
  gather_draws(mu_ps[t]) %>%
  median_qi() %>%
  mutate(year = years[t]+63)

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)+
  geom_line(data = res_pred, aes(year, .value)) +
  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "2nd order random walk fit and forecast shown in black")
```



Question 3

Run the first order and second order random walk models, including projections out 10 years. Compare these estimates with the linear fit by plotting everything on the same graph.

- Answer:

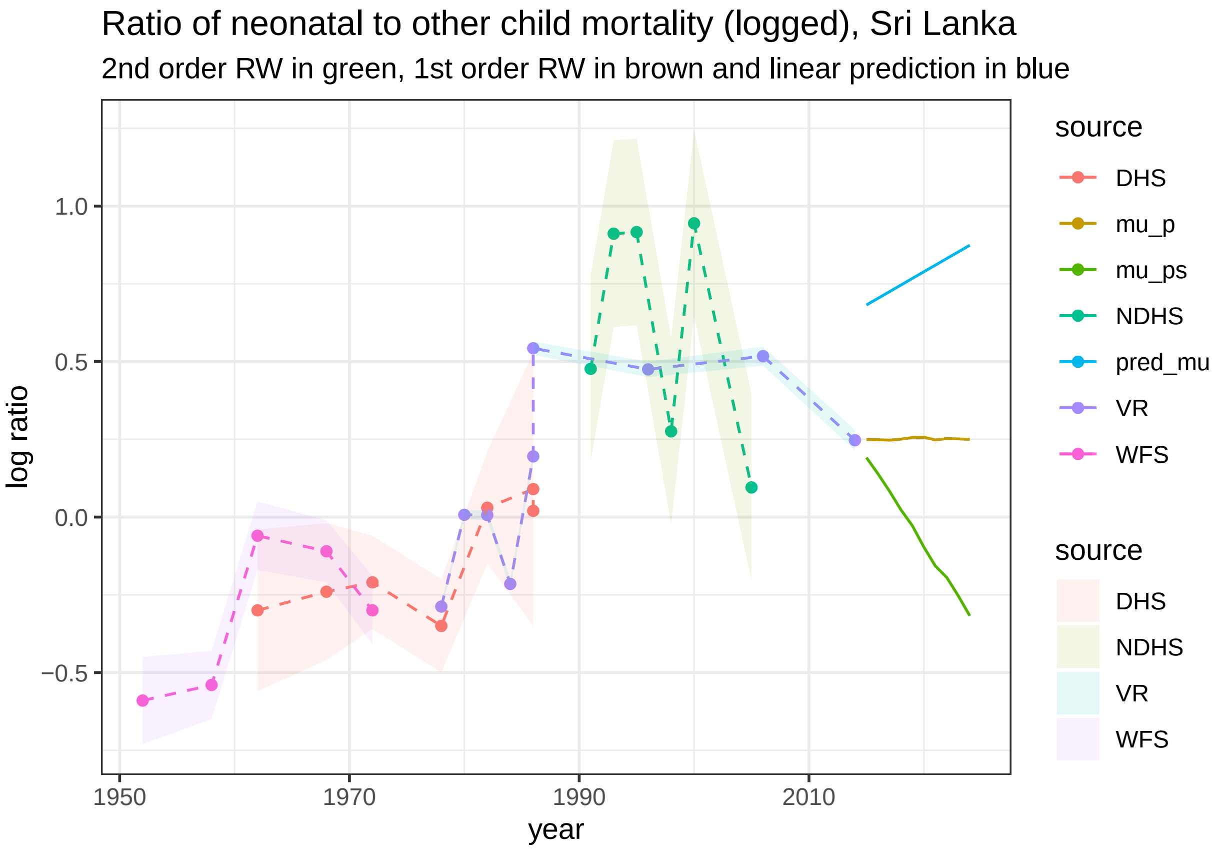
By just looking at three different estimations. We could see the second order random walk gave the most reasonable estimation. The first order random walk follows and the linear model prediction is the worst. Which aligns with our intuition.

```
Q4_fit <- stan(file = "./lka_rw_me.stan", data = Q3data)
```

```
## Trying to compile a simple C file
```

```
res_pred_first <- Q4_fit %>%
  gather_draws(mu_p[t]) %>%
  median_qi() %>%
  mutate(year = years[t]+63)

res_pred_second <- Q3_fit %>%
  gather_draws(mu_ps[t]) %>%
  median_qi() %>%
  mutate(year = years[t]+63)

pred <- res_pred_first[,c(2,3,9)] %>%
  rbind(res_pred_second[,c(2,3,9)]) %>%
  rbind(res_pred_linear[,c(2,3,9)])
```

```
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 = pred, aes(year, .value, color = .variable)) +
  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "2nd order RW in green, 1st order RW in brown and linear prediction
```



Question 4

Rerun the RW2 model excluding the VR data. Briefly comment on the differences between the two data situations.

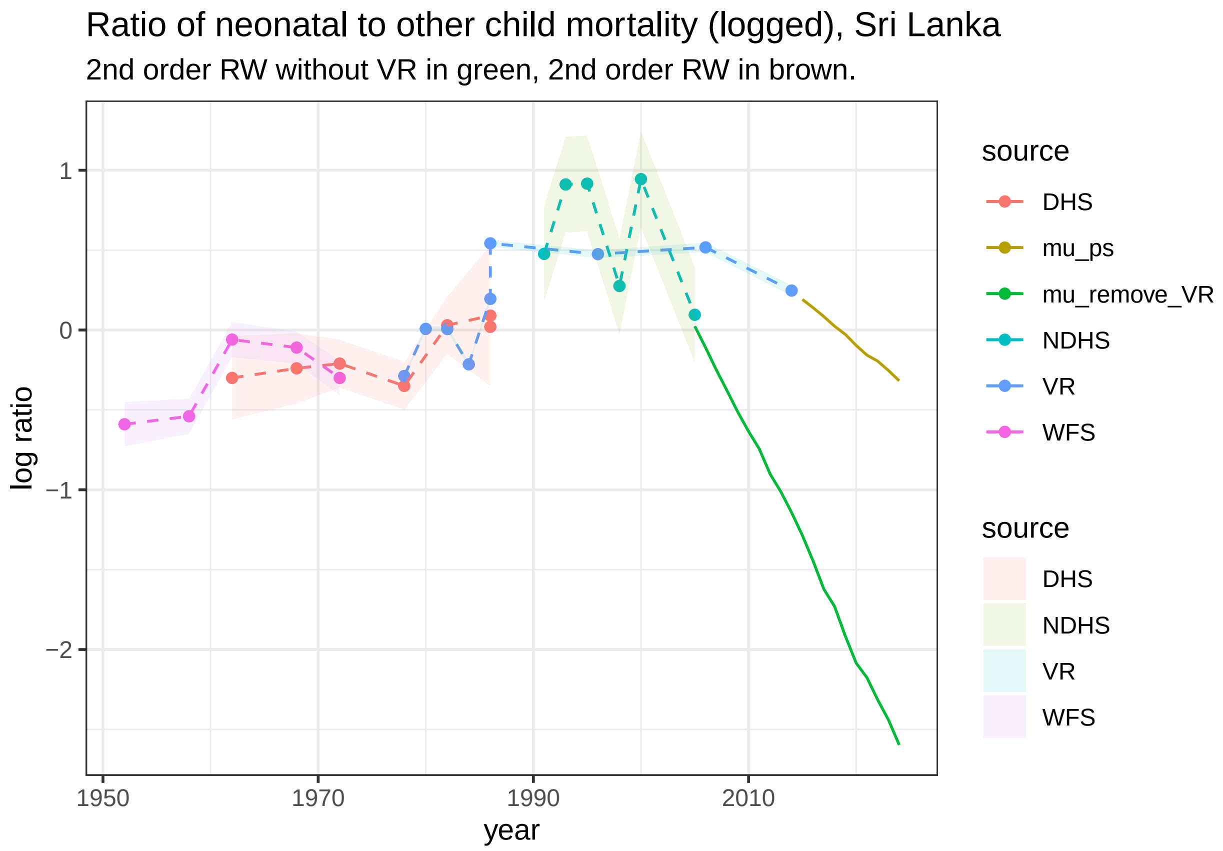
- Answer:
- First since we removed the source belong to VR, the maximum year in our sample will decrease from 2014 to 2005. So we need to change our start year for prediction for comparisons.
- Second since we have removed the VR variable. Our prediction after will most rely on NDHS. So we could see the prediction without VR will decrease much faster than the prediction with VR.

```
lka_Q4 = lka[lka$source != "VR", ]
observed_years <- lka_Q4$year
years <- min(observed_years):max(observed_years)
```

```
nyears <- length(years)

Q4data <- list(y = lka_Q4$logit_ratio,
               year_i = observed_years - years[1]+1,
               T = nyears, years = years,
               N = length(observed_years),
               P = 20,
               se = lka_Q4$se)

Q4_fit <- stan(file = "./lka_rw_second.stan", data = Q4data)
```

```
## Warning: There were 1 chains where the estimated Bayesian Fraction of Missing Information was low. Se
## http://mc-stan.org/misc/warnings.html#bfmi-low

## Warning: Examine the pairs() plot to diagnose sampling problems

## Warning: Bulk Effective Samples Size (ESS) is too low, indicating posterior means and medians may be
## Running the chains for more iterations may help. See
## http://mc-stan.org/misc/warnings.html#bulk-ess

## Warning: Tail Effective Samples Size (ESS) is too low, indicating posterior variances and tail quant
## Running the chains for more iterations may help. See
## http://mc-stan.org/misc/warnings.html#tail-ess
```

```
res_pred_second_Q4 <- Q4_fit %>%
  gather_draws(mu_ps[t]) %>%
  median_qi() %>%
  mutate(year = years[t]+53)

res_pred_second_Q4$.variable = "mu_remove_VR"

pred <- rbind(res_pred_second[,c(2,3,9)]) %>%
  rbind(res_pred_second_Q4[,c(2,3,9)])

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 = pred, aes(year, .value, color = .variable)) +
  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "2nd order RW without VR in green, 2nd order RW in brown.")
```



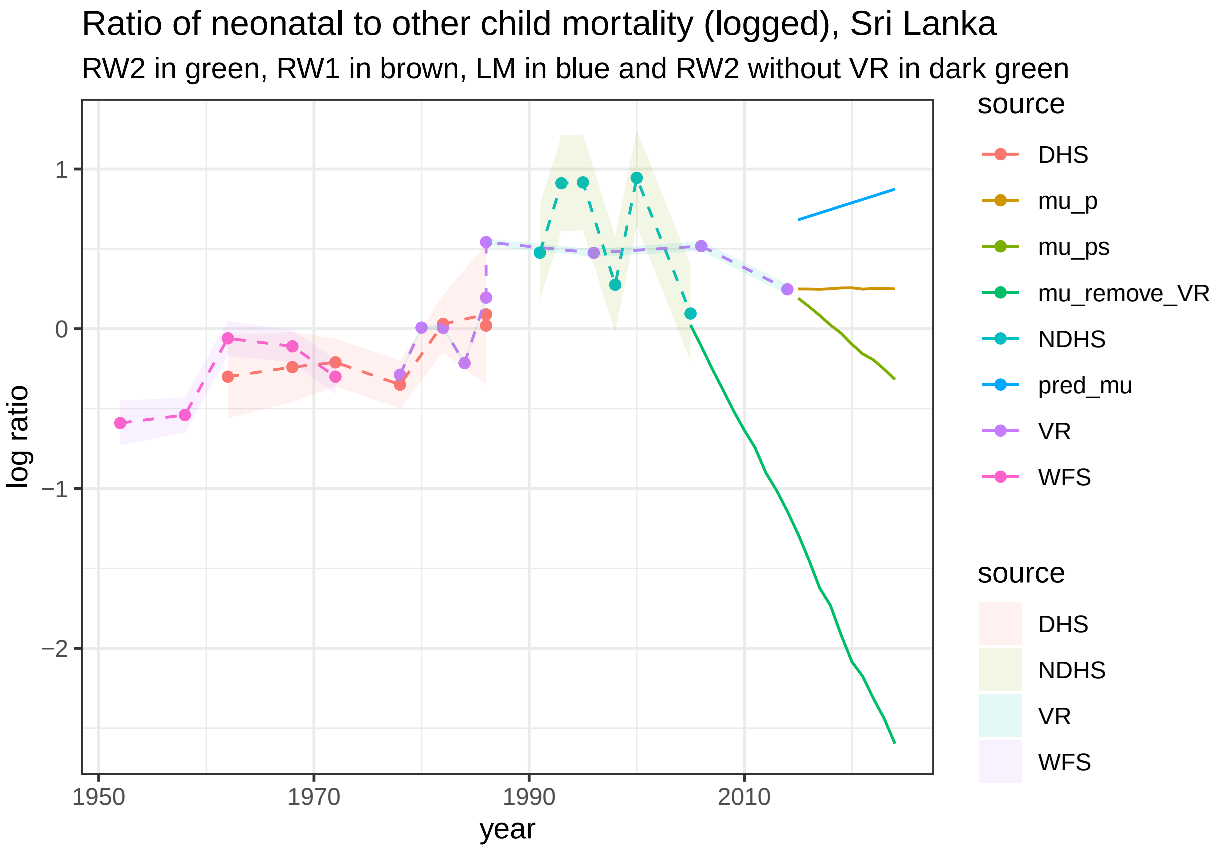
Question 5

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

- Answer:

From the plot we could see the the RW2 is the most appropriate. It includes all the information and reflect the most recent trend. The linear model is obviously not very accurate. The trend of RW1 model is too weak and the trend of RW2 without VR is not informative due to the lack of data. I guess arima could be a better fit.

```
pred <- pred <- res_pred_first[,c(2,3,9)] %>%
  rbind(res_pred_second[,c(2,3,9)]) %>%
  rbind(res_pred_linear[,c(2,3,9)]) %>%
  rbind(res_pred_second_Q4[,c(2,3,9)])

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 = pred, aes(year, .value, color = .variable)) +
  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "RW2 in green, RW1 in brown, LM in blue and RW2 without VR in dark g
```


Ratio of neonatal to other child mortality (logged), Sri Lanka
RW2 in green, RW1 in brown, LM in blue and RW2 without VR in dark green
log ratio
1
0
−1
−2
1950
1970
1990
2010
year
source
DHS
mu_p
mu_ps
mu_remove_VR
NDHS
pred_mu
VR
WFS
source
DHS
NDHS
VR
WFS