Sta2201 Lab 10

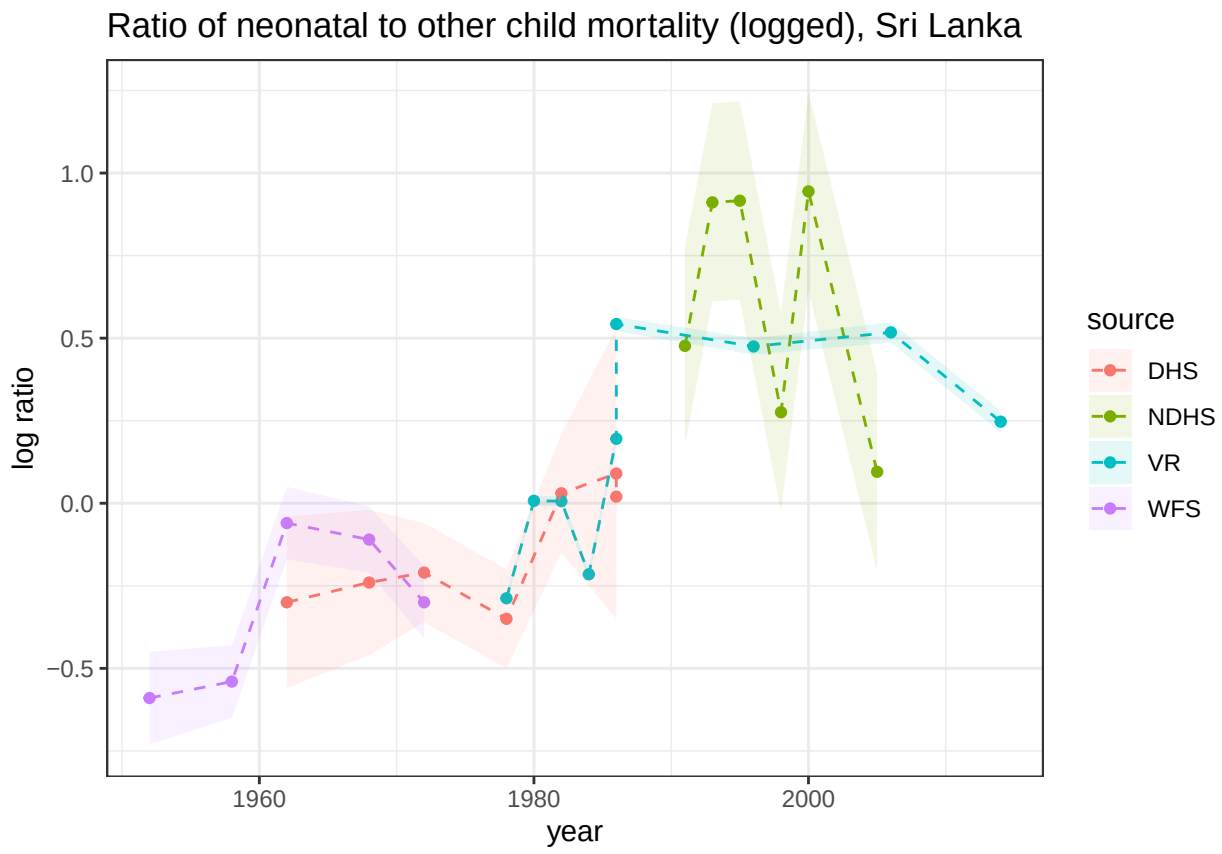
Yeonjoon Choi

2023-03-23

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("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 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 = "linear_mod.stan", refresh = 0)
```

Extract the results:

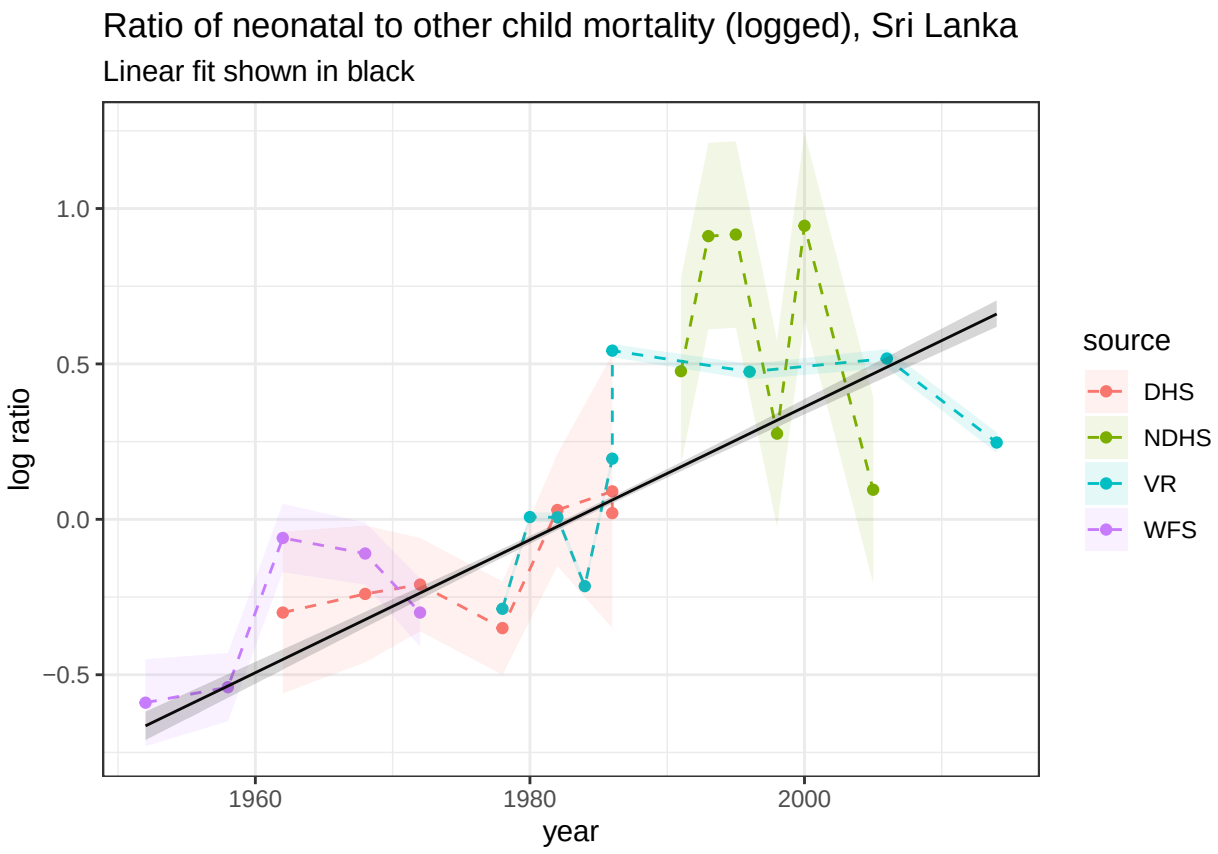
```
res <- mod %>%
  gather_draws(mu[t]) %>%
  median_qi() %>%
  mutate(year = years[t])

res
```

```
## # A tibble: 63 x 9
```

```
##        t .variable .value .lower .upper .width .point .interval  year
##    <int> <chr>      <dbl>  <dbl>  <dbl>  <dbl> <chr>  <chr>     <int>
##  1     1 mu        -0.665 -0.710 -0.619   0.95 median qi         1952
##  2     2 mu        -0.643 -0.688 -0.599   0.95 median qi         1953
##  3     3 mu        -0.622 -0.665 -0.579   0.95 median qi         1954
##  4     4 mu        -0.600 -0.642 -0.559   0.95 median qi         1955
##  5     5 mu        -0.579 -0.620 -0.539   0.95 median qi         1956
##  6     6 mu        -0.558 -0.597 -0.519   0.95 median qi         1957
##  7     7 mu        -0.536 -0.574 -0.499   0.95 median qi         1958
##  8     8 mu        -0.515 -0.552 -0.479   0.95 median qi         1959
##  9     9 mu        -0.494 -0.529 -0.459   0.95 median qi         1960
## 10    10 mu        -0.472 -0.506 -0.439   0.95 median qi         1961
## # ... with 53 more rows
```

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")
```



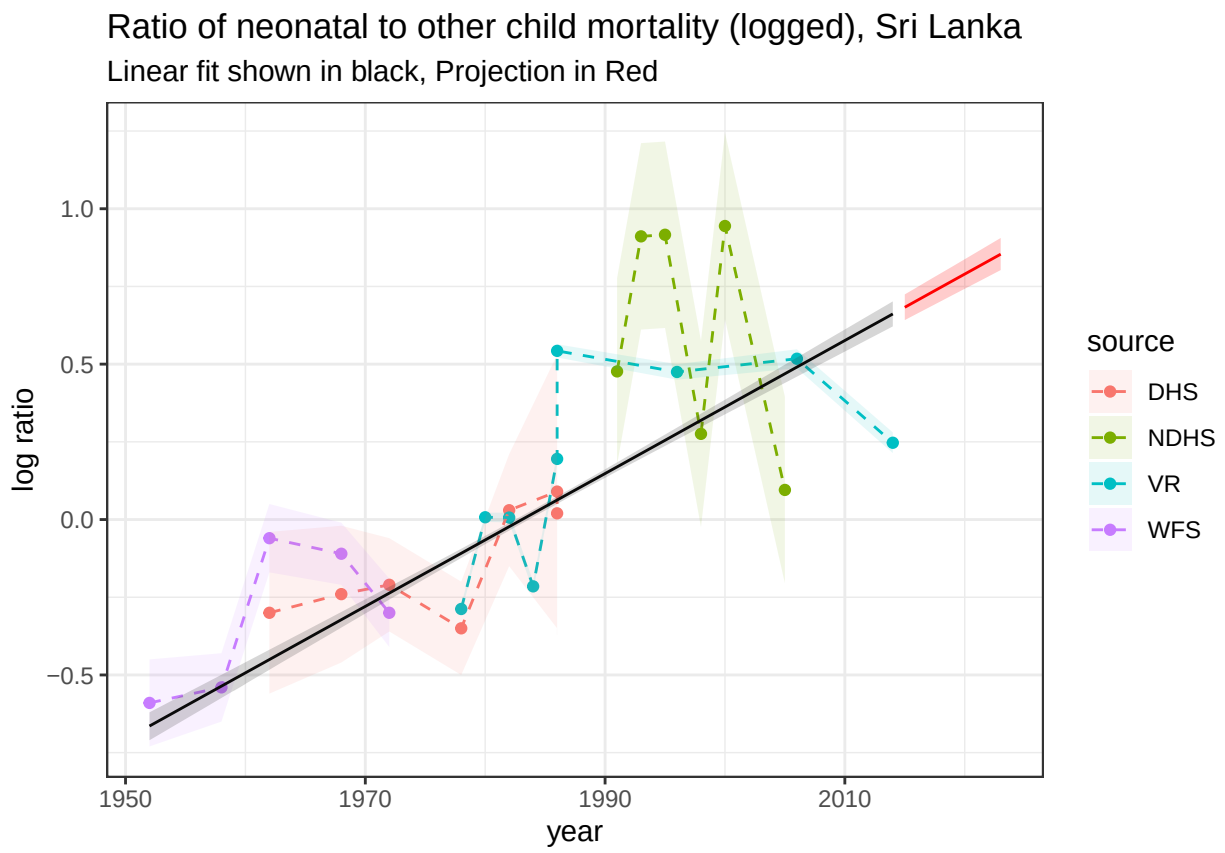
Question 1

Project the linear model above out to 2023 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.

```
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, P=9)
mod2 <- stan(data = stan_data,
             file = "lienar_mod_with_projection.stan", refresh = 0)
```

```
res = mod2 |>
  gather_draws(mu[t])|>
  median_qi()|>
  mutate(year = years[t])


res_p = mod2|>
  gather_draws(mu_p[p])|>
  median_qi()|>
  mutate(year = years[nyears]+p)
```

```
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_p, aes(year, .value), col = "red") +
  geom_ribbon(data = res_p, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2, fill = "red")+
  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "Linear fit shown in black, Projection in Red")
```



Random walks

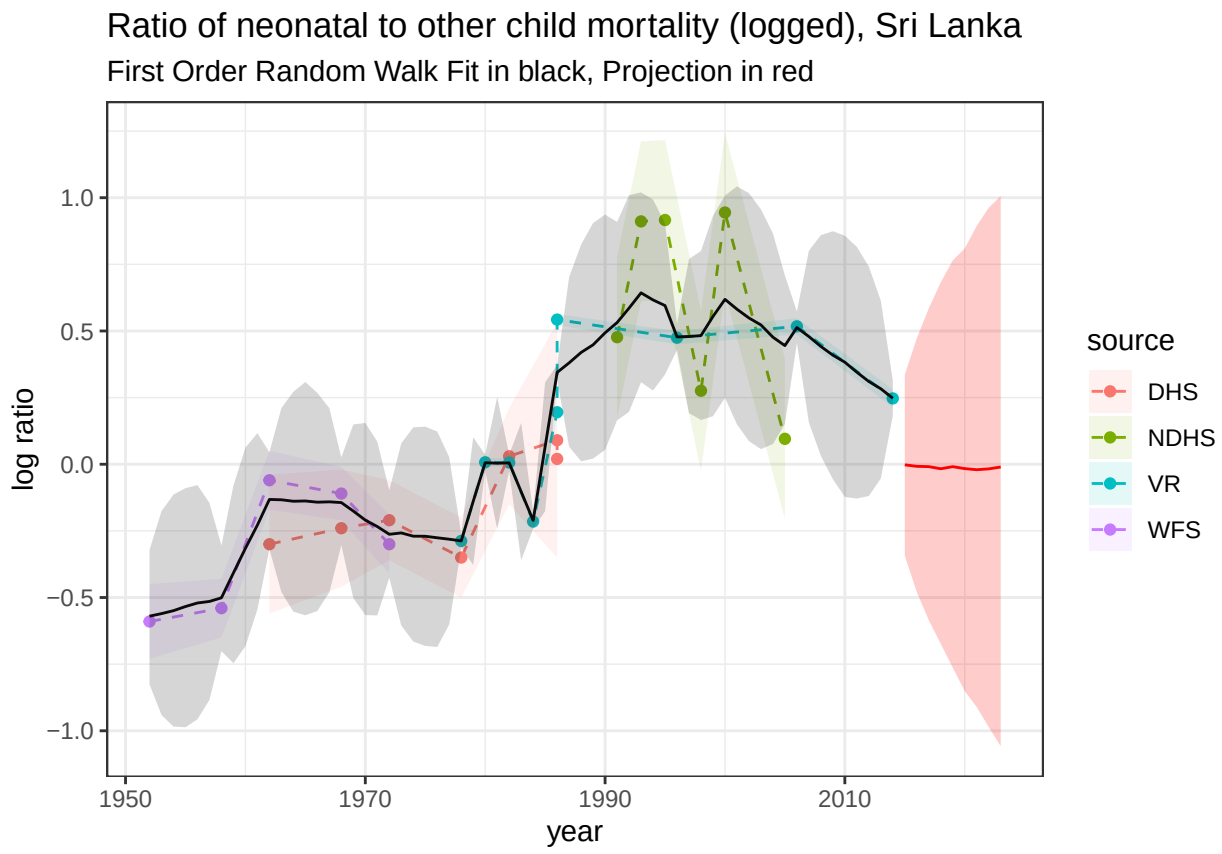
Question 2

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 2023.

```
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, P=9)
mod3 <- stan(data = stan_data,
             file = "random_walk_mod.stan", refresh = 0)
```

```
res = mod3 |>
  gather_draws(mu[t])|>
  median_qi()|>
  mutate(year = years[t])


res_p = mod3|>
  gather_draws(mu_p[p])|>
  median_qi()|>
  mutate(year = years[nyears]+p)
```

```
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_p, aes(year, .value), col = "red") +
  geom_ribbon(data = res_p, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2, fill = "red")+
  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "First Order Random Walk Fit in black, Projection in red")
```



Question 3

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

```
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, P=9)
mod4 <- stan(data = stan_data,
             file = "2nd_order_random_walk.stan", refresh = 0)
```

Fit and Projection

```
res = mod4 |>
  gather_draws(mu[t])|>
  median_qi()|>
  mutate(year = years[t])


res_p = mod4|>
  gather_draws(mu_p[p])|>
  median_qi()|>
```

```
  mutate(year = years[nyears]+p)

res
```

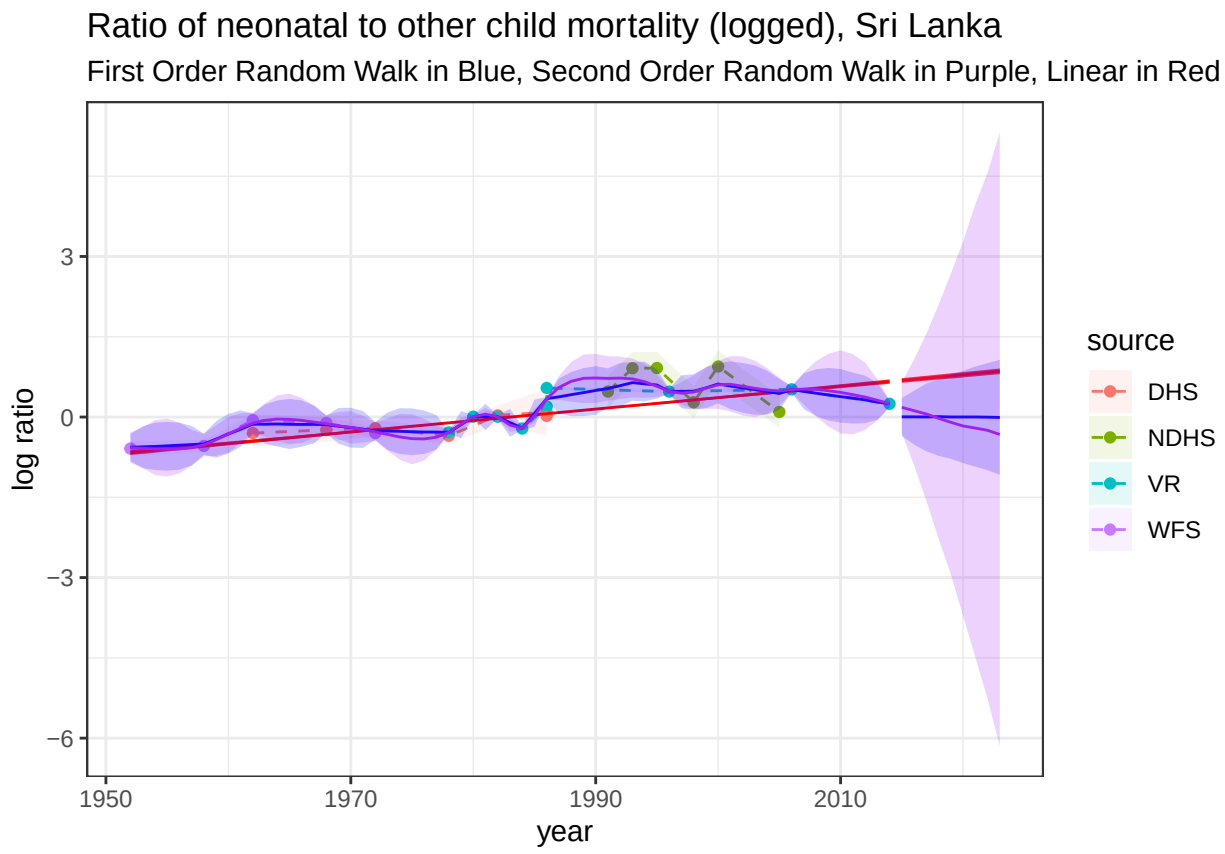
```
## # A tibble: 63 x 9
##        t .variable .value .lower   .upper .width .point .interval  year
##    <int> <chr>      <dbl>  <dbl>    <dbl>  <dbl> <chr>  <chr>     <int>
##  1     1 mu        -0.578 -0.858 -0.299     0.95 median qi         1952
##  2     2 mu        -0.596 -0.977 -0.185     0.95 median qi         1953
##  3     3 mu        -0.609 -1.12  -0.0866    0.95 median qi         1954
##  4     4 mu        -0.617 -1.16  -0.0589    0.95 median qi         1955
##  5     5 mu        -0.608 -1.10  -0.118     0.95 median qi         1956
##  6     6 mu        -0.576 -0.921 -0.235     0.95 median qi         1957
##  7     7 mu        -0.520 -0.736 -0.309     0.95 median qi         1958
##  8     8 mu        -0.425 -0.717 -0.145     0.95 median qi         1959
##  9     9 mu        -0.317 -0.637  0.00271   0.95 median qi         1960
## 10    10 mu        -0.203 -0.469  0.0628    0.95 median qi         1961
## # ... with 53 more rows
```

```
res_p
```

```
## # A tibble: 9 x 9
##       p .variable   .value .lower .upper .width .point .interval  year
##   <int> <chr>        <dbl>  <dbl>  <dbl>  <dbl> <chr>  <chr>     <int>
## 1     1 mu_p       0.184   -0.295  0.641   0.95 median qi         2015
## 2     2 mu_p       0.119   -0.886  1.07    0.95 median qi         2016
## 3     3 mu_p       0.0515  -1.50   1.58    0.95 median qi         2017
## 4     4 mu_p      -0.00745 -2.19   2.12    0.95 median qi         2018
## 5     5 mu_p      -0.0564  -2.94   2.72    0.95 median qi         2019
## 6     6 mu_p      -0.109   -3.78   3.37    0.95 median qi         2020
## 7     7 mu_p      -0.162   -4.60   4.00    0.95 median qi         2021
## 8     8 mu_p      -0.216   -5.50   4.68    0.95 median qi         2022
## 9     9 mu_p      -0.268   -6.41   5.43    0.95 median qi         2023
```

Question 4

Run the first order and second order random walk models, including projections out to 2023. Compare these estimates with the linear fit by plotting everything on the same graph.

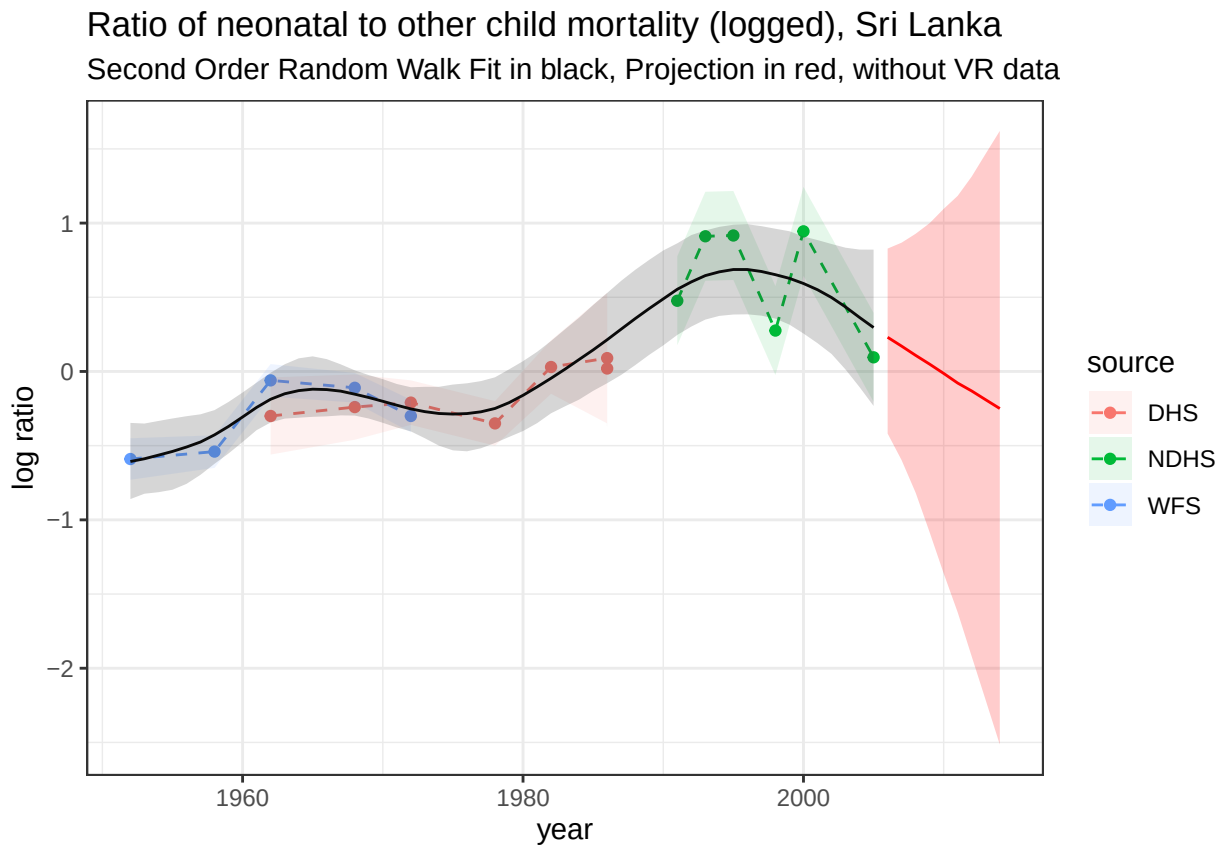


Question 5

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

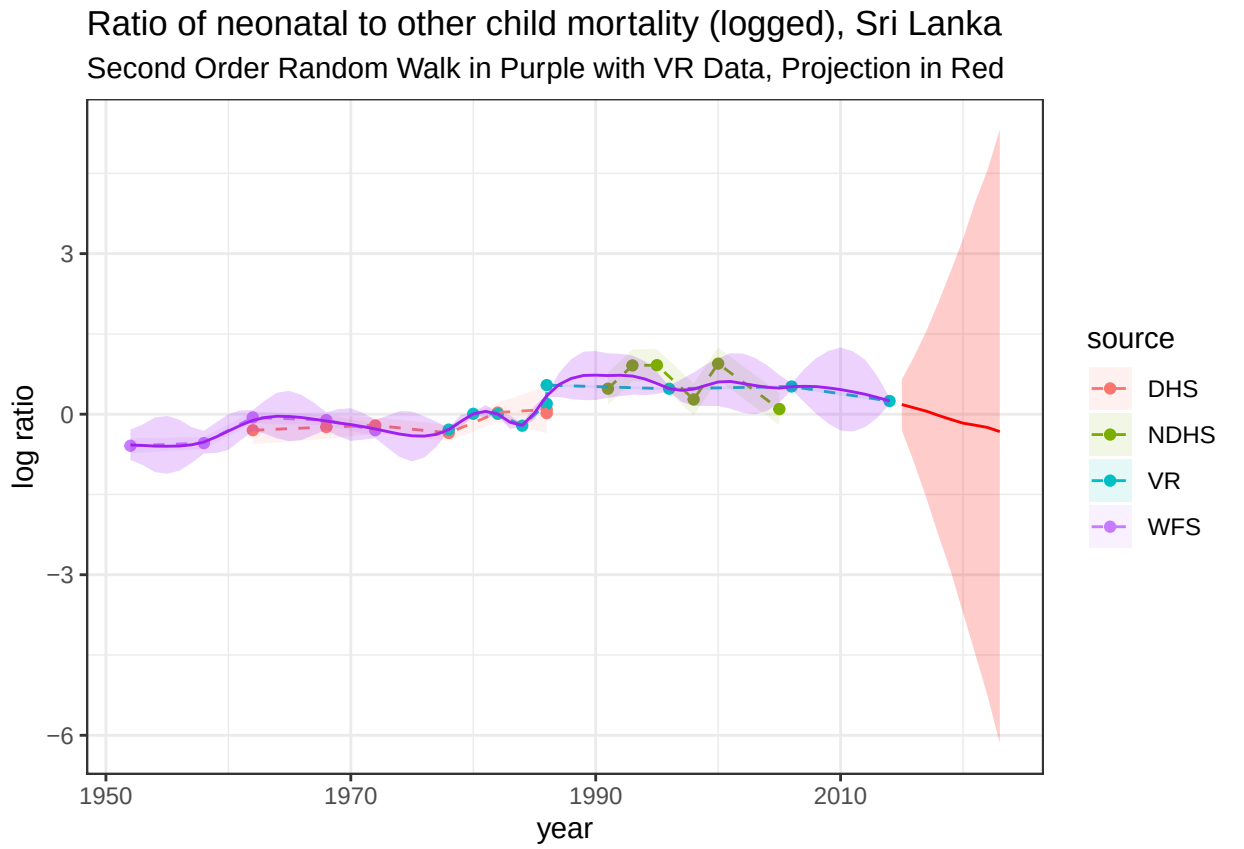
```
new_data = lka|>
  filter(source != "VR")


observed_years <- new_data$year
years <- min(observed_years):max(observed_years)
nyears <- length(years)
stan_data <- list(y = new_data$logit_ratio, year_i = observed_years - years[1]+1,
                  T = nyears, years = years, N = length(observed_years),
                   se = new_data$se, P=9)
mod5 <- stan(data = stan_data,
             file = "2nd_order_random_walk.stan", refresh = 0)


res = mod5 |>
  gather_draws(mu[t])|>
  median_qi()|>
  mutate(year = years[t])
```

```
res_p = mod5|>
  gather_draws(mu_p[p])|>
  median_qi()|>
  mutate(year = years[nyears]+p)


ggplot(new_data, 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_p, aes(year, .value), col = "red") +
  geom_ribbon(data = res_p, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2, fill = "red")+
  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "Second Order Random Walk Fit in black, Projection in red, without VR
```




```
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_4, aes(year, .value), col= "purple") +
geom_ribbon(data = res_4, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2, fill = "purple")
  geom_line(data = res_p_4, aes(year, .value), col = "red") +
geom_ribbon(data = res_p_4, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2, fill = "red")+
theme_bw()+
labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
     y = "log ratio", subtitle = "Second Order Random Walk in Purple with VR Data, Projection in Red")
```



Note that VR has noticeably small standard error compared to other data set. Hence, in the fit with VR, the time period with VR has small credible interval, representing smaller variability in the fit. Without VR data, the fit has larger credible interval for the years that VR data set covered.

Question 6

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

The second order random walk produced reasonable fit, but the projection has too much variability. Hence, hierarchical model structure on σ for different sources with second order random walk may produce better fit.