Week 10: Temporal data

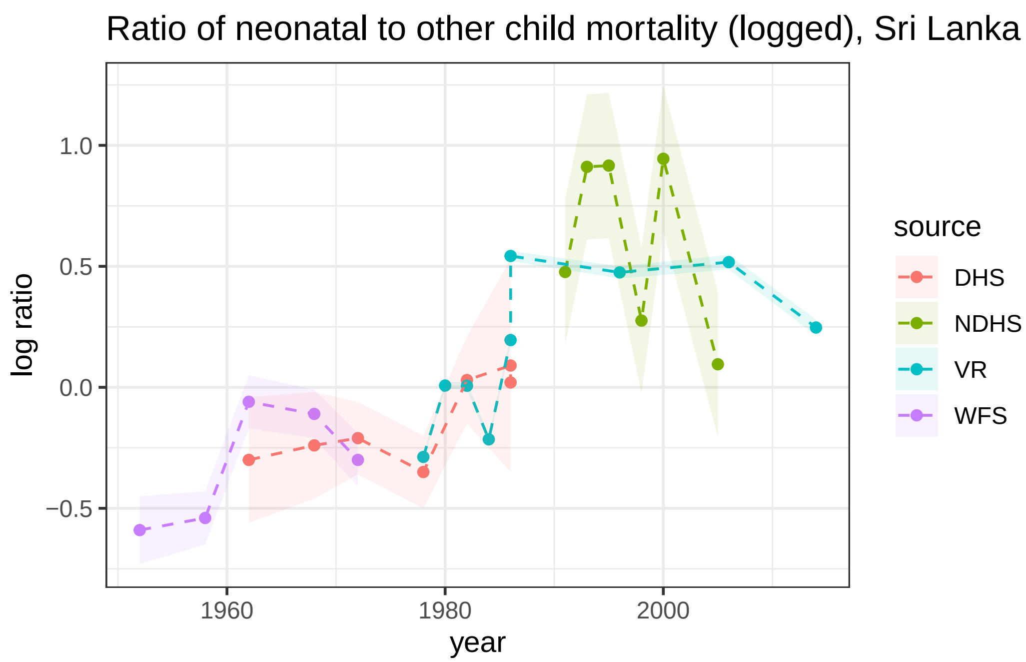
Jin Xin HU

24/03/23

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 = "lka_linear_me.stan",refresh = 0)
```

Extract the results:

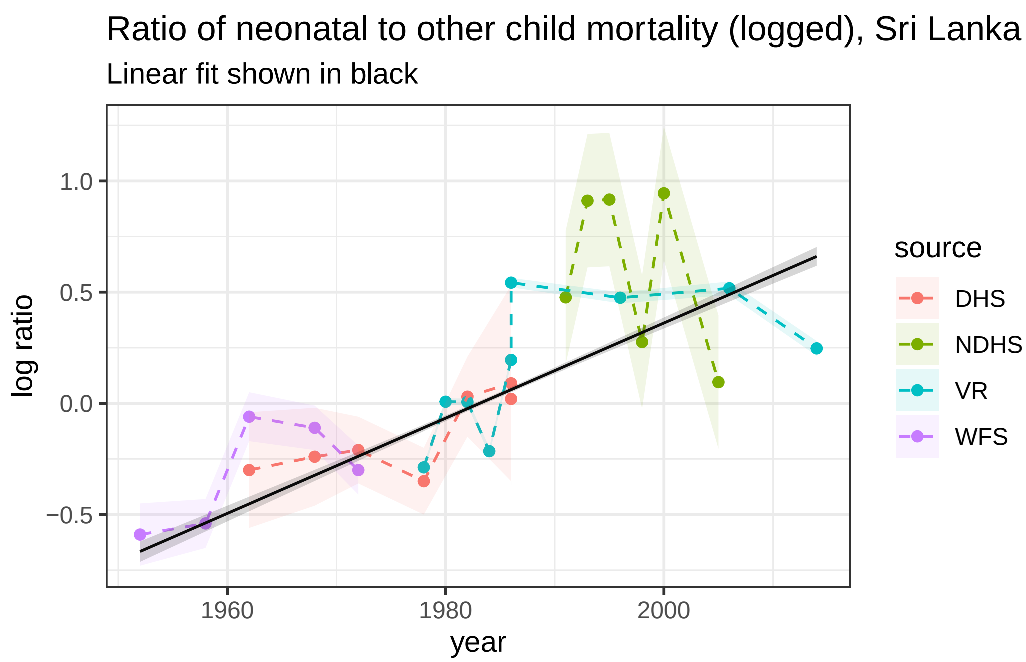
```
res <- mod %>%
  gather_draws(mu[t]) %>%
  median_qi() %>%
  mutate(year = years[t])
```

```r
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        -0.666 -0.713 -0.621   0.95 median qi         1952
 2     2 mu        -0.645 -0.690 -0.601   0.95 median qi         1953
 3     3 mu        -0.623 -0.667 -0.581   0.95 median qi         1954
 4     4 mu        -0.602 -0.644 -0.561   0.95 median qi         1955
 5     5 mu        -0.581 -0.622 -0.540   0.95 median qi         1956
 6     6 mu        -0.559 -0.599 -0.520   0.95 median qi         1957
 7     7 mu        -0.538 -0.576 -0.500   0.95 median qi         1958
 8     8 mu        -0.516 -0.553 -0.480   0.95 median qi         1959
 9     9 mu        -0.495 -0.531 -0.460   0.95 median qi         1960
10    10 mu        -0.473 -0.508 -0.440   0.95 median qi         1961
# ... with 53 more rows
```

Plot the results:

```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, 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

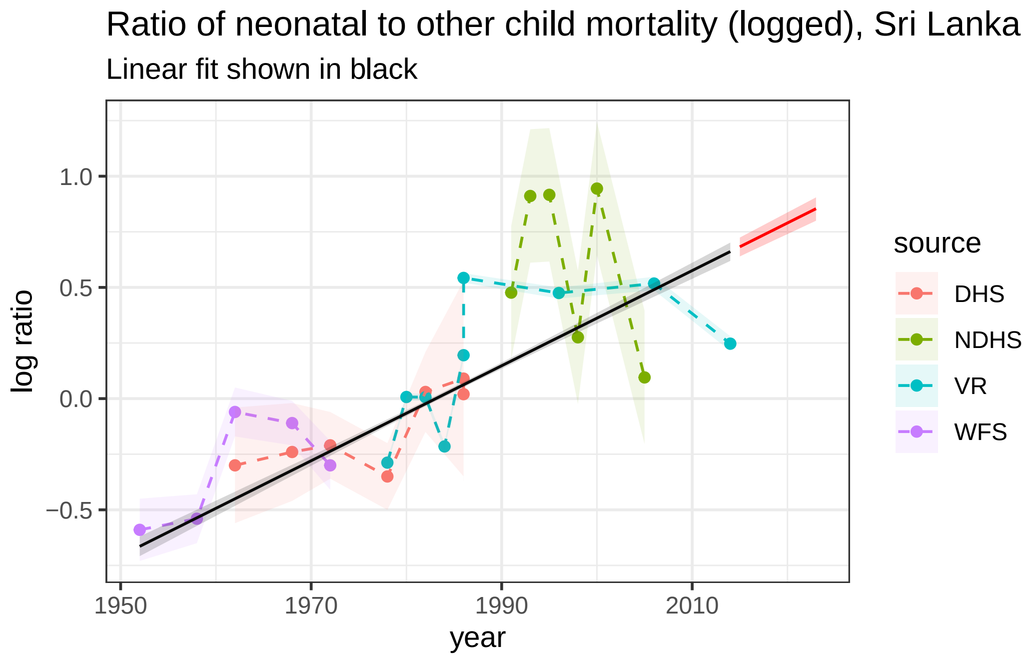
Project the linear model above out to 2023 by adding a `generated quantities` block in Stan (do the projections based on the expected value μ). Plot the resulting projections on a graph similar to that above.

```
P=2023-max(observed_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=P)

mod1 <- stan(data = stan_data,
             file = "lab10_lka_linear.stan",
             refresh = 0)
```

```
#linear model
res1 <- mod1 |>
  gather_draws(mu[t]) |>
  median_qi() |>
  mutate(year = years[t],model = " The linear")
```

```
res1_p <- mod1 |>
  gather_draws(mu_p[p]) |>
  median_qi() |>
  mutate(year = years[nyears]+p,model = " The linear")
```

```
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_p, aes(year, .value), col='red') +
  geom_ribbon(data = res1_p, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2, f
  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "Linear fit shown in black")
```



Random walks

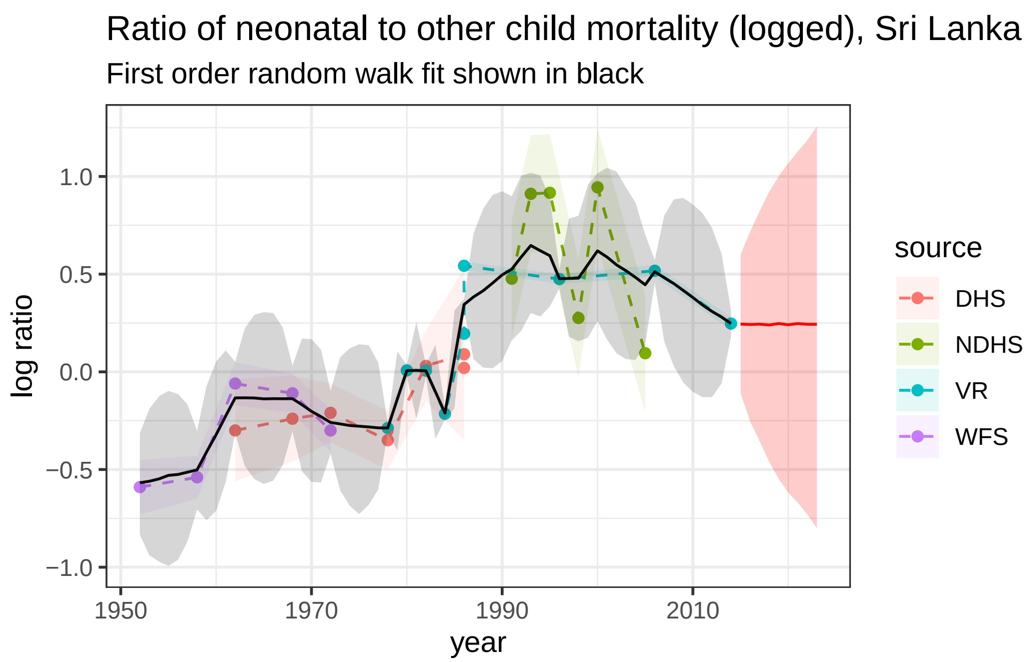
Question 2

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.

```
mod2 <- stan(data = stan_data,
             file = "lab10_lka_RW1.stan",
             refresh = 0)
```

```
#1st Order RW  model
res2 <- mod2 |>
  gather_draws(mu[t]) |>
  median_qi() |>
  mutate(year = years[t],model = "the first Order RW")

res2_p <- mod2 |>
  gather_draws(mu_p[p]) |>
  median_qi() |>
  mutate(year = years[nyears]+p,model = "the first Order RW")
```

```
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 = res2, aes(year, .value)) +
  geom_ribbon(data = res2, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2)+
  geom_line(data = res2_p, aes(year, .value), col='red') +
  geom_ribbon(data = res2_p, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2, f
  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "First order random walk fit shown in black")
```



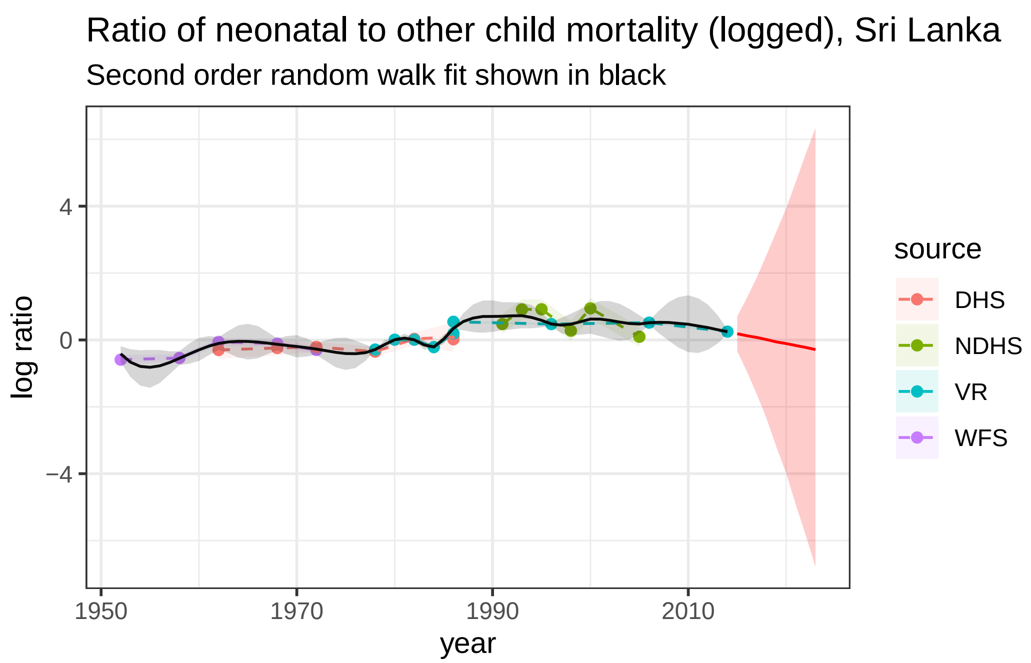
Question 3

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

```
mod3 <- stan(data = stan_data,
             file = "lab10_lka_RW2.stan",
             refresh = 0)
```

```
# the second Order RW model
res3 <- mod3 |>
  gather_draws(mu[t]) |>
  median_qi() |>
  mutate(year = years[t],model = "the second Order RW" )

res3_p <- mod3 |>
  gather_draws(mu_p[p]) |>
  median_qi() |>
  mutate(year = years[nyears]+p, model = "the second Order RW")
```

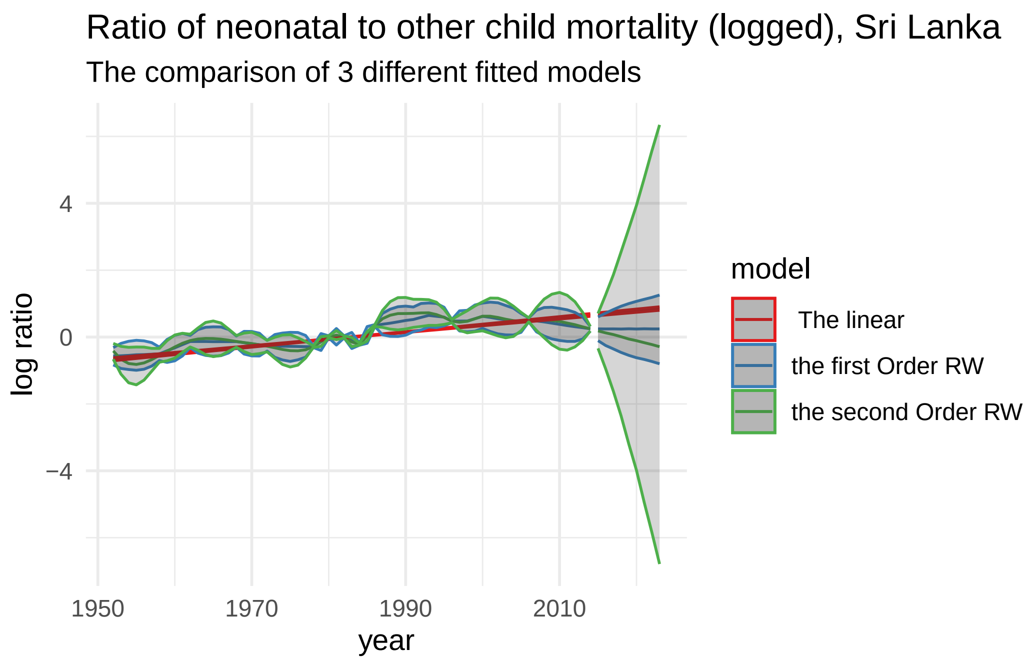
```r
ggplot(lka, aes(year, logit_ratio)) +
  geom_point(aes( color = source)) +
  geom_line(aes( color = source), lty = 2) +
  geom_ribbon(aes(ymin = logit_ratio - se,
                  ymax = logit_ratio + se,
                  fill =  source), alpha = 0.1) +
  theme_bw()+
  geom_line(data = res3, aes(year, .value)) +
  geom_ribbon(data = res3, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2)+
  geom_line(data = res3_p, aes(year, .value), col='red') +
  geom_ribbon(data = res3_p, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2, f
  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "Second order random walk fit shown in black")
```



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.

```r
res_total <- rbind(res1, res2, res3)
res_p_total <- rbind(res1_p, res2_p, res3_p)

ggplot(res_total, aes(x=year, y=.value, color = model)) +
  geom_line() +
  geom_ribbon(aes(y = .value, ymin = .lower, ymax = .upper, color = model), alpha = 0.2)+
  geom_line(data = res_p_total, aes(x=year, y=.value, color = model)) +
  geom_ribbon(data = res_p_total, aes(y = .value, ymin = .lower, ymax = .upper, color = mo
  theme_minimal()+scale_color_brewer(palette="Set1")+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "The comparison of 3 different fitted models")
```



Question 5

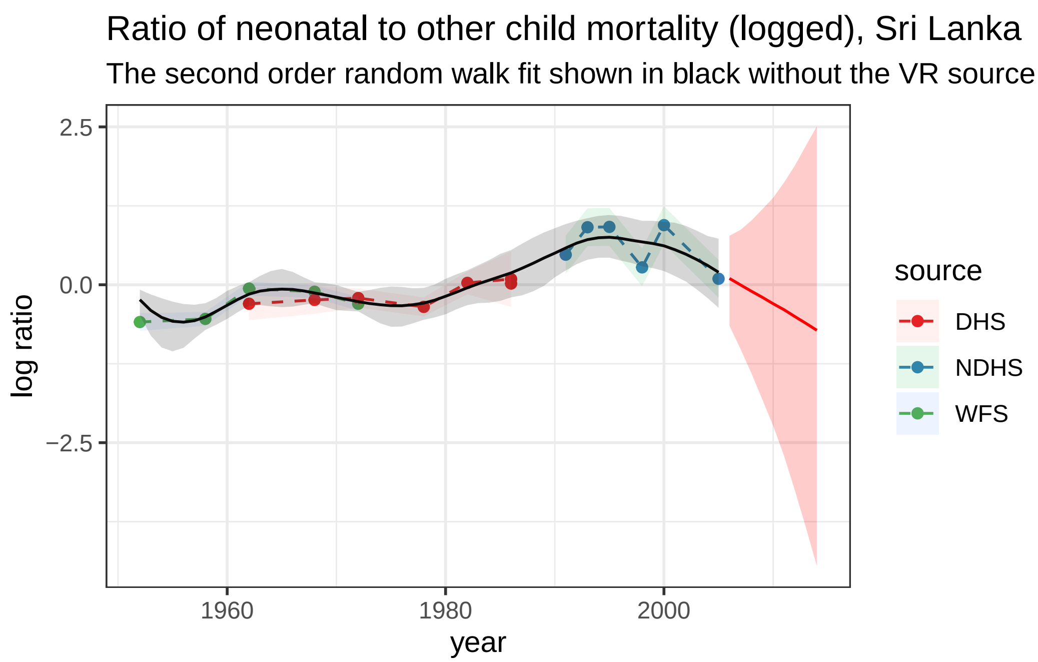
Rerun the RW2 model excluding the VR data. Briefly comment on the differences between the two data situations.

Here VR data means data from vital registration systems which is the best and reliable source.

```
no_vr_lka <- lka |>filter(source != "VR")

no_vr_observed_years <- no_vr_lka$year
no_vr_years <- min(no_vr_observed_years):max(no_vr_observed_years)
no_vr_nyears <- length(no_vr_years)
N <- length(no_vr_observed_years)

no_vr_stan_data <- list(y = no_vr_lka$logit_ratio, year_i = no_vr_observed_years - no_vr_y
                        T = no_vr_nyears, years = no_vr_years, N = N, se = no_vr_lka$se, P
```

```
mod4 <- stan(data = no_vr_stan_data,
             file = "lab10_lka_RW2.stan",
             refresh = 0)


res4 <- mod4 |>
  gather_draws(mu[t]) |>
  median_qi() |>
  mutate(year = no_vr_years[t])

res4_p <- mod4 |>
  gather_draws(mu_p[p]) |>
  median_qi() |>
  mutate(year = no_vr_years[no_vr_nyears]+p)
```

```
ggplot(no_vr_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 = res4, aes(year, .value)) +
  geom_ribbon(data = res4, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2)+
  geom_line(data = res4_p, aes(year, .value), col='red') +
  geom_ribbon(data = res4_p, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2, f
  theme_bw()+scale_color_brewer(palette="Set1")+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "The second order random walk fit shown in black withou
```



From the above graph we could see that the uncertainty or CI is more stable among whole data set than the figure of question 4. It is reduced gradually to project. The exclusion of VR data means that some reliable extra data points were missed. Nevertheless, the uncertainty from the projecting is still decreased because deleting VR data may diminished significantly the variance among the different sources, which would lead to the shrinkage of its projection.

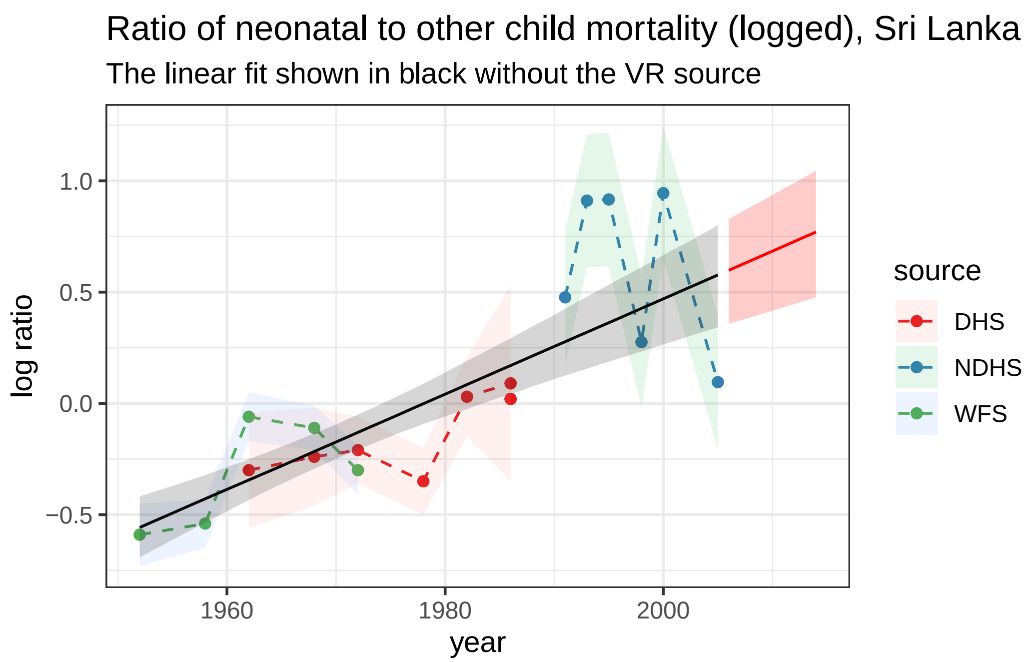
Question 6

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

For linear model without VR sources:

```
no_vr_lka <- lka |>filter(source != "VR")

no_vr_observed_years <- no_vr_lka$year
no_vr_years <- min(no_vr_observed_years):max(no_vr_observed_years)
no_vr_nyears <- length(no_vr_years)
N <- length(no_vr_observed_years)
m_year<- as.integer(mean(no_vr_years))
```

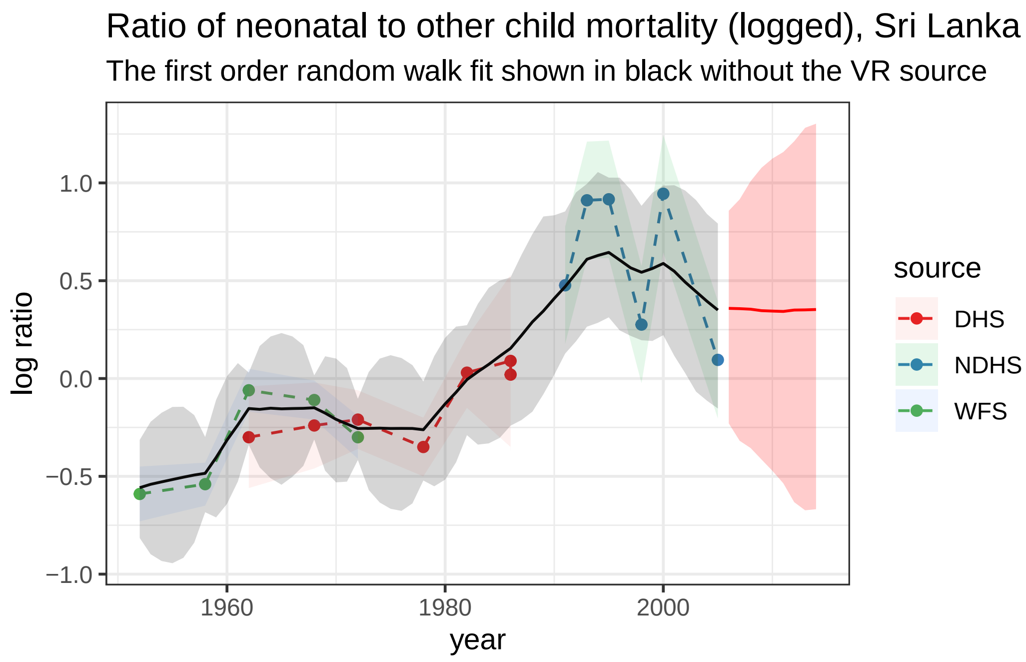
```r
no_vr_stan_data <- list(y = no_vr_lka$logit_ratio, year_i = no_vr_observed_years - no_vr_y
                        T = no_vr_nyears, years = no_vr_years, N = N,
                        mid_year = m_year, se = no_vr_lka$se, P=P)


mod5 <- stan(data = no_vr_stan_data,
             file = "lab10_lka_linear.stan",
             refresh = 0)

res5 <- mod5 |>
  gather_draws(mu[t]) |>
  median_qi() |>
  mutate(year = no_vr_years[t])

res5_p <- mod5 |>
  gather_draws(mu_p[p]) |>
  median_qi() |>
  mutate(year = no_vr_years[no_vr_nyears]+p)

ggplot(no_vr_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 = res5, aes(year, .value)) +
  geom_ribbon(data = res5, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2)+
  geom_line(data = res5_p, aes(year, .value), col='red') +
  geom_ribbon(data = res5_p, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2, f
  theme_bw()+scale_color_brewer(palette="Set1")+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "The linear fit shown in black without the VR source")
```



For first order random walk model without VR sources:

```
no_vr_lka <- lka |>filter(source != "VR")

no_vr_observed_years <- no_vr_lka$year
no_vr_years <- min(no_vr_observed_years):max(no_vr_observed_years)
no_vr_nyears <- length(no_vr_years)
N <- length(no_vr_observed_years)

no_vr_stan_data <- list(y = no_vr_lka$logit_ratio, year_i = no_vr_observed_years - no_vr_y
                        T = no_vr_nyears, years = no_vr_years, N = N, se = no_vr_lka$se, P

mod6 <- stan(data = no_vr_stan_data,
             file = "lab10_lka_RW1.stan",
             refresh = 0)

res6 <- mod6 |>
  gather_draws(mu[t]) |>
  median_qi() |>
  mutate(year = no_vr_years[t])

res6_p <- mod6 |>
```

```r
  gather_draws(mu_p[p]) |>
  median_qi() |>
  mutate(year = no_vr_years[no_vr_nyears]+p)


ggplot(no_vr_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 = res6, aes(year, .value)) +
  geom_ribbon(data = res6, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2)+
  geom_line(data = res6_p, aes(year, .value), col='red') +
  geom_ribbon(data = res6_p, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2, f
  theme_bw()+scale_color_brewer(palette="Set1")+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "The first order random walk fit shown in black without
```



I think that the three models have their cons and pros although the three models still were good:

1) the linear model is too simple and easy to be used, but a little vague,thus there is no exact prediction.

2) 1st and 2nd random walk models could show promise in picking up characteristics of time series but useless for understanding why changes are happening, and whether they are likely to happen in future.

I just choose an alternative model-Bayesian hierarchical state-space model.The reason is that:

1) It could hold a whole suite of candidate covariates.

2) It could have a association between interest outcomes and covariates are allowed to vary by geography and over time (in a smooth way).For example,we can put a time series model on the regression coefficients.

3) It is useful in understanding changes in observed outcomes as well, in a regression framework and better projection.