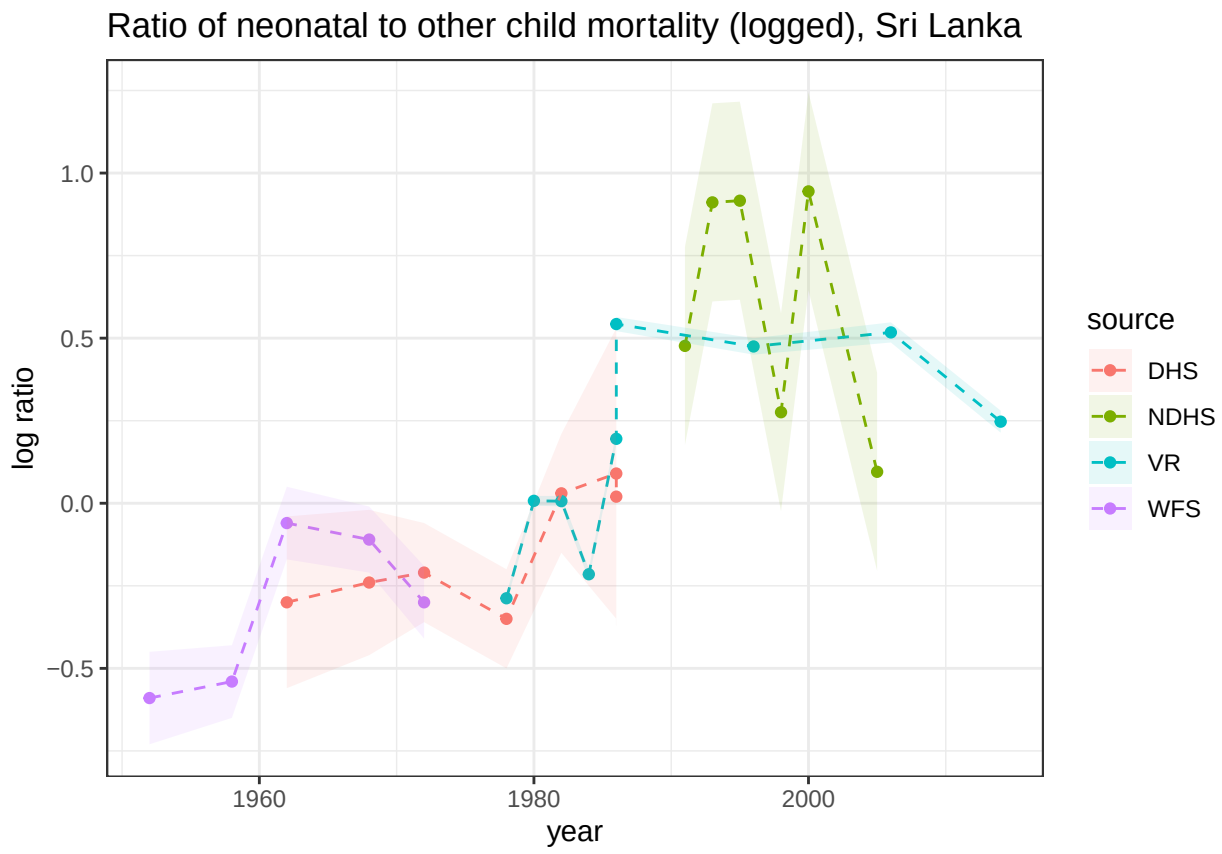
Week 10: Temporal data

Qiaoyu Liang

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:

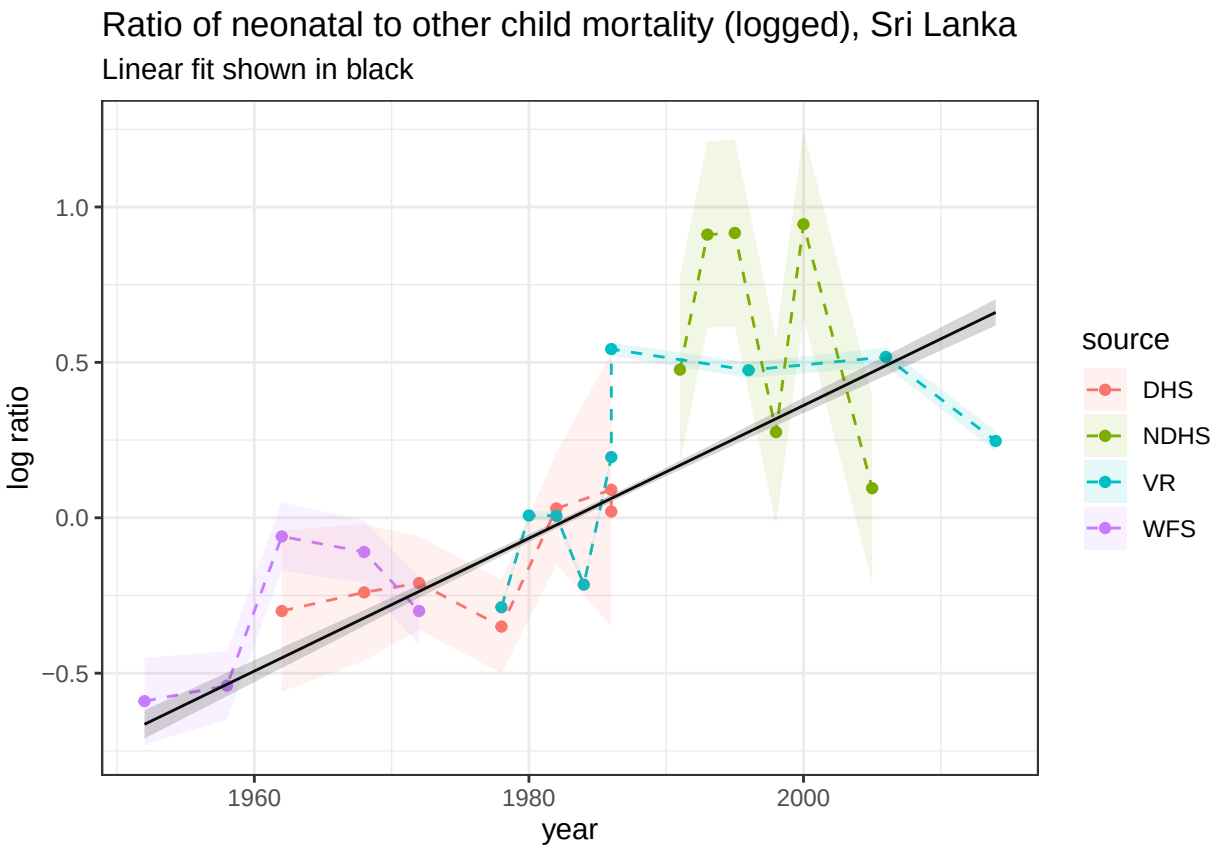
```
observed_years <- lka$year
years <- min(observed_years):max(observed_years)
nyears <- length(years)
stan_data <- list(y = lka$logit_ratio, year_i = observed_years - years[1]+1,
                  T = nyears, years = years, N = length(observed_years),
                  mid_year = mean(years), se = lka$se)
mod <- stan(data = stan_data,
            file = "lka_linear_me.stan",
            seed = 2201)
```

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

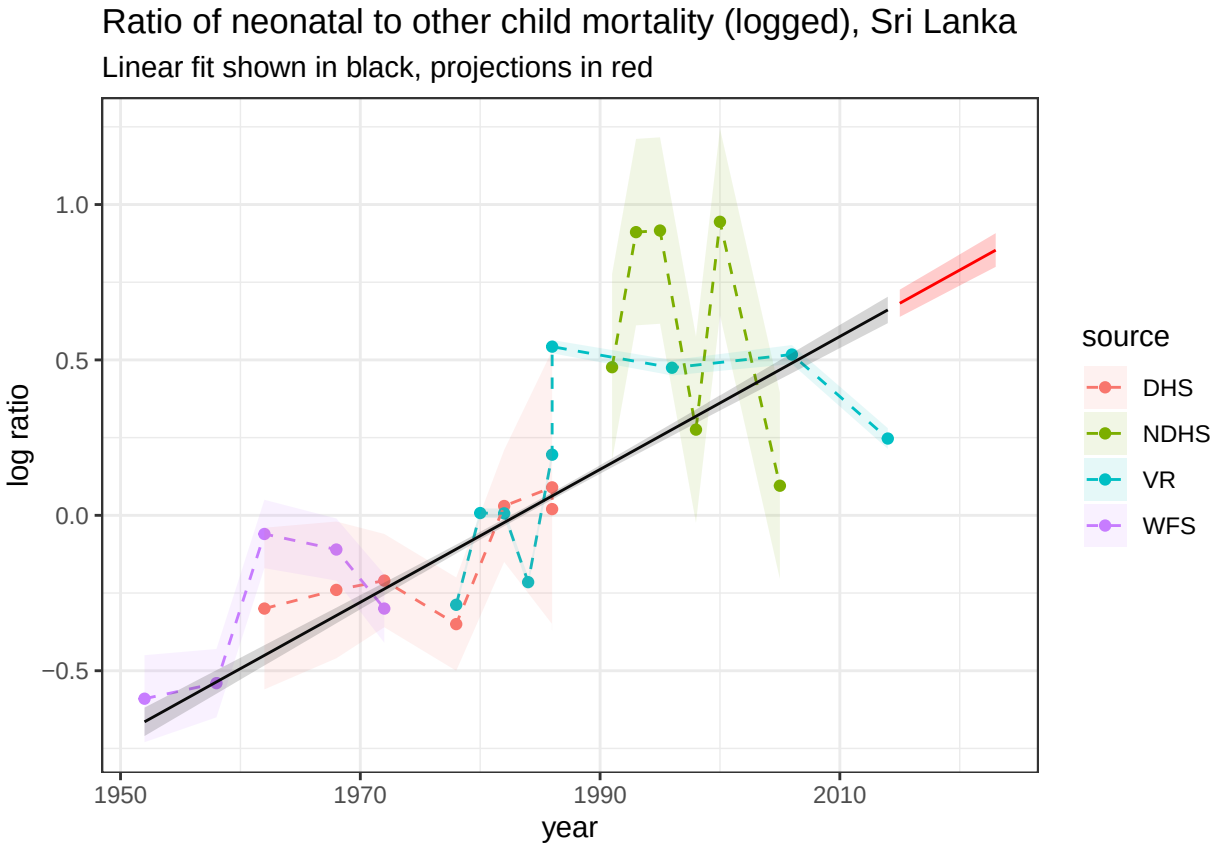
Project the linear model above out to 2023 by adding a `generated quantities` block in Stan (do the projections based on the expected value μ). Plot the resulting projections on a graph similar to that above.

```
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)
modQ1 <- stan(data = stan_data,
              file = "Lab10_mod1.stan",
              seed = 2201)
```

```
res1 = modQ1 %>%
  gather_draws(mu[t]) %>%
  median_qi() %>%
  mutate(year=years[t])
res1_p = modQ1 %>%
  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 = 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),
              fill="red", 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, projections 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.

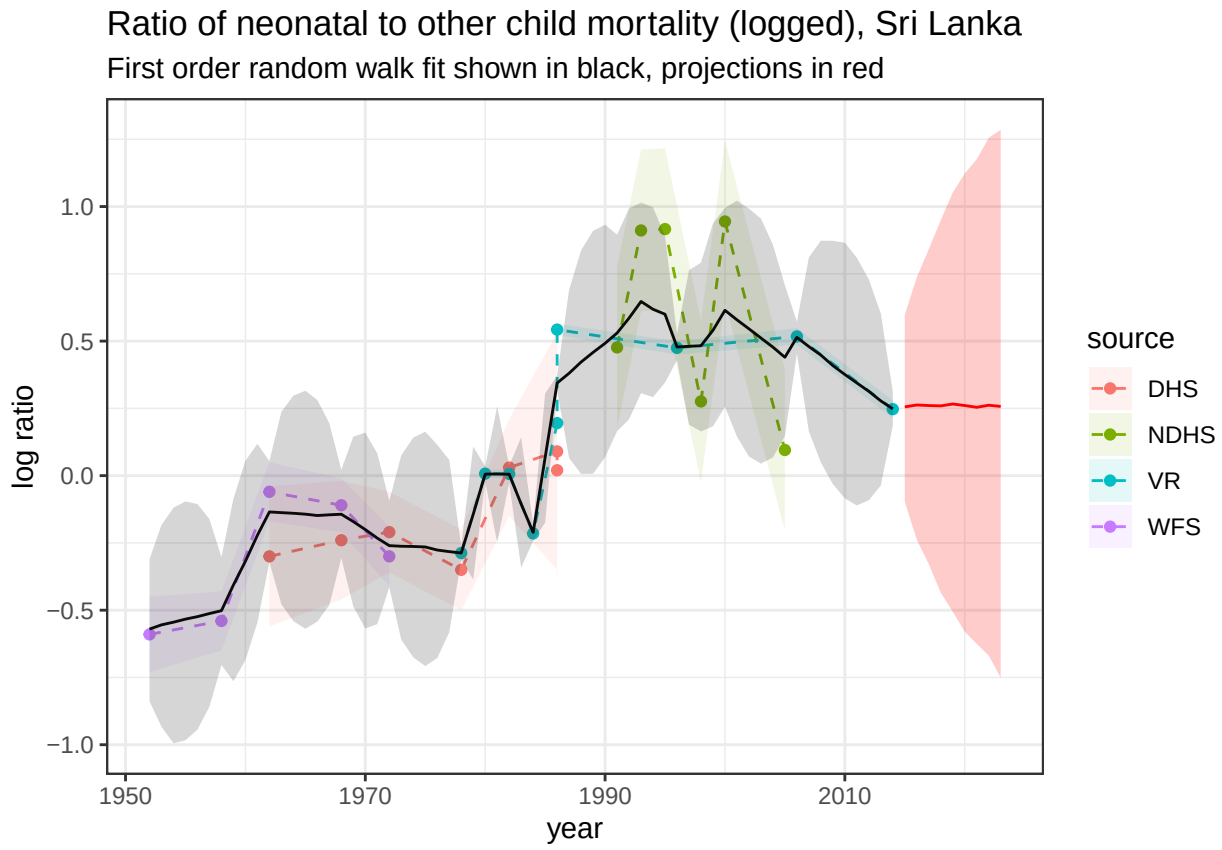
```
modQ2 <- stan(data = stan_data,
              file = "Lab10_mod2.stan",
              seed = 2201)
```

```
res2 <- modQ2 %>%
  gather_draws(mu[t])  %>%
  median_qi()  %>%
  mutate(year = years[t])
res2_p <- modQ2  %>%
  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 = 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),
              fill="red", alpha = 0.2)+

  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, projections in red")
```



Question 3

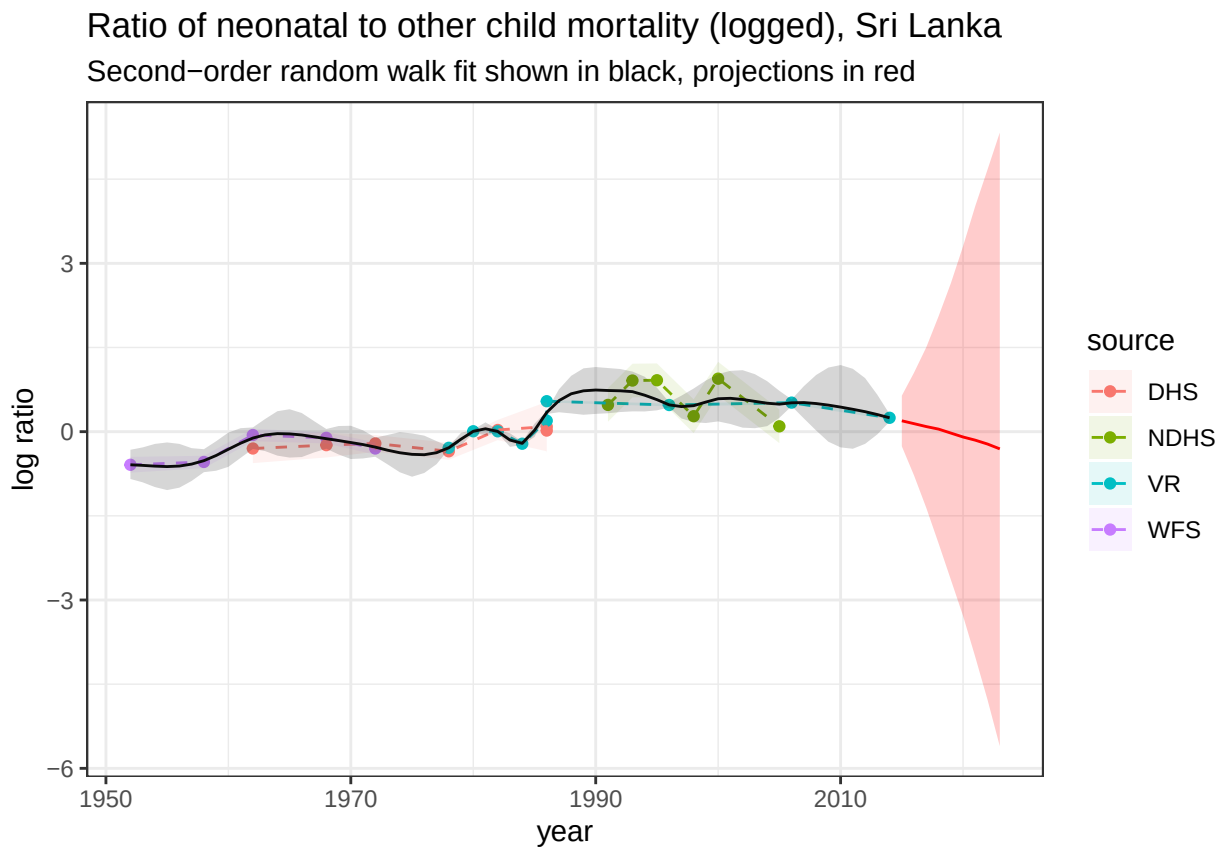
Now alter your model above to estimate and project a second-order random walk model (RW2).

```
modQ3 <- stan(data = stan_data,
              file = "Lab10_mod3.stan",
              seed = 2201)
```

```
res3 <- modQ3 %>%
  gather_draws(mu[t]) %>%
  median_qi() %>%
  mutate(year = years[t])
res3_p <- modQ3 %>%
  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 = 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),
              fill="red", alpha = 0.2)+

  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, projections in red")
```



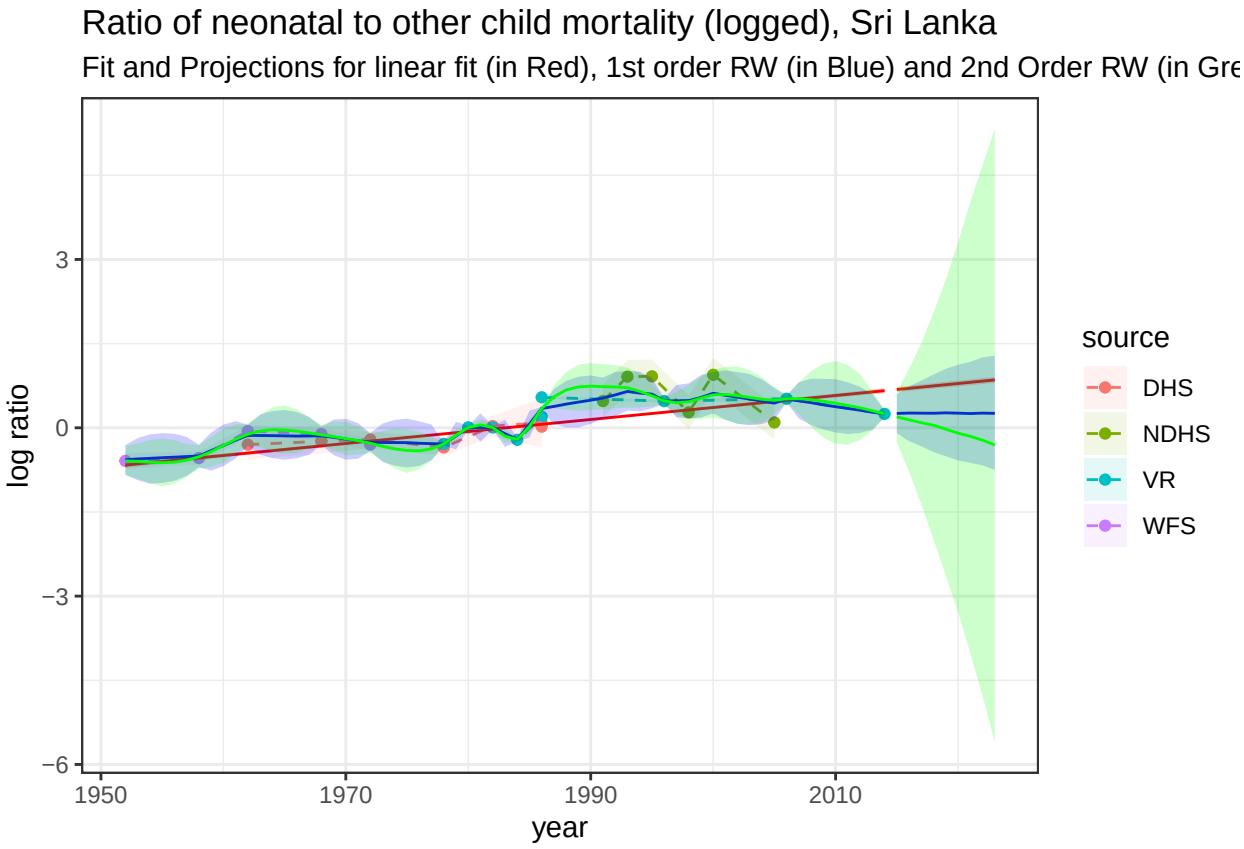
Question 4

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.

```
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),col="red") +
  geom_ribbon(data = res1, aes(y = .value, ymin = .lower, ymax = .upper),
              fill="red", 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),
              fill="red", alpha = 0.2)+

  geom_line(data = res2, aes(year, .value),col="blue") +
  geom_ribbon(data = res2, aes(y = .value, ymin = .lower, ymax = .upper),
              fill="blue", alpha = 0.2)+
  geom_line(data = res2_p, aes(year, .value),col="blue") +
  geom_ribbon(data = res2_p, aes(y = .value, ymin = .lower, ymax = .upper),
              fill="blue", alpha = 0.2)+
```

```
  geom_line(data = res3, aes(year, .value),col="green") +
  geom_ribbon(data = res3, aes(y = .value, ymin = .lower, ymax = .upper),
              fill="green", alpha = 0.2)+
  geom_line(data = res3_p, aes(year, .value),col="green") +
  geom_ribbon(data = res3_p, aes(y = .value, ymin = .lower, ymax = .upper),
              fill="green", alpha = 0.2)+
  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio",
       subtitle = "Fit and Projections for linear fit (in Red), 1st order RW (in Blue) and 2nd Order RW
```



Question 5

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

```
lka_no_vr <- lka %>%
  filter(source != "VR")
observed_years <- lka_no_vr$year
years <- min(observed_years):max(observed_years)
nyears <- length(years)
stan_data <- list(y = lka_no_vr$logit_ratio, year_i = observed_years - years[1]+1,
                  T = nyears, years = years, N = length(observed_years),
```

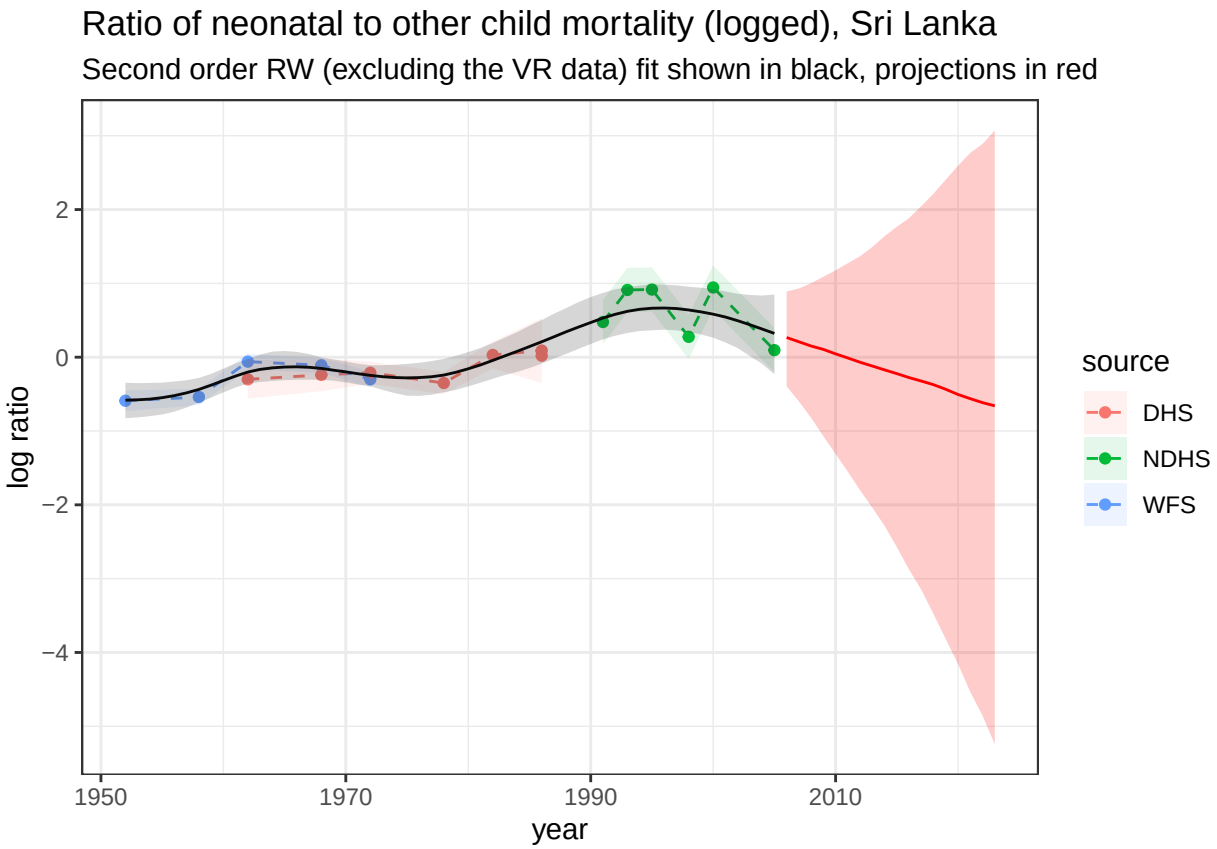
```
                  mid_year = mean(years),
                  se = lka_no_vr$se, P=18)
```

```
modQ5 <- stan(data = stan_data,
              file = "Lab10_mod3.stan",
              seed = 2201)
```

```
res4 <- modQ5 %>%
  gather_draws(mu[t]) %>%
  median_qi() %>%
  mutate(year = years[t])
res4_p <- modQ5 %>%
  gather_draws(mu_p[p]) %>%
  median_qi() %>%
  mutate(year = years[nyears]+p)
```

```
ggplot(lka_no_vr, 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),
              fill="red", alpha = 0.2)+

  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio",
       subtitle = "Second order RW (excluding the VR data) fit shown in black, projections in red")
```



Comment: From the above plot, we can notice that without the VR data, the 2nd order random walk model's uncertainty is much smoother than the 2nd order random walk model's uncertainty with VR data. We also notice that we only have data up until 2005 instead of 2014 if we do not have VR data.

Question 6

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

Comment: I think the most appropriate model is second order random walk model with VR data since it has reasonable point estimates and seems to give valid projections. The projections from second order random walk model with VR data appear to be more plausible than those from the linear model, where the log ratio of neonatal mortality is projected to constantly increase, and the 1st order random walk, where the log ratio is projected as being more or less constant. Furthermore, I do not think we should exclude the VR data, unless we know the data is not reliable. The reason for not excluding the VR data is that the VR data provides additional information for the years from 2005 to 2014 and appears to strongly influence the projections.