Week 10: Temporal data

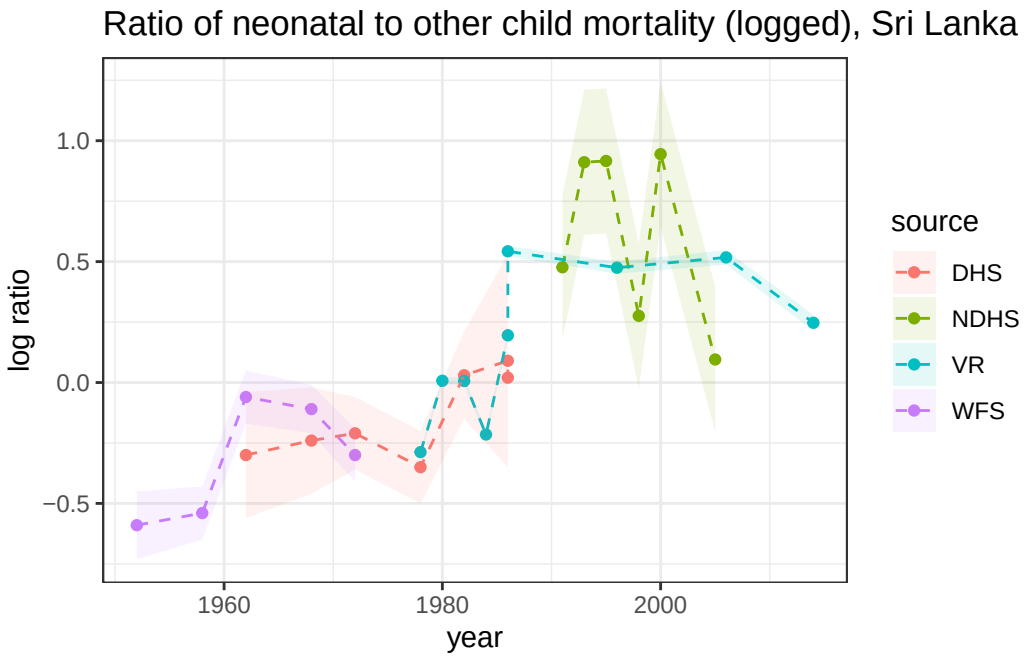
J. Arturo Esquivel

22/03/23

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(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("lka_linear_me.stan"),
            refresh = 0,
            verbose = FALSE)
```

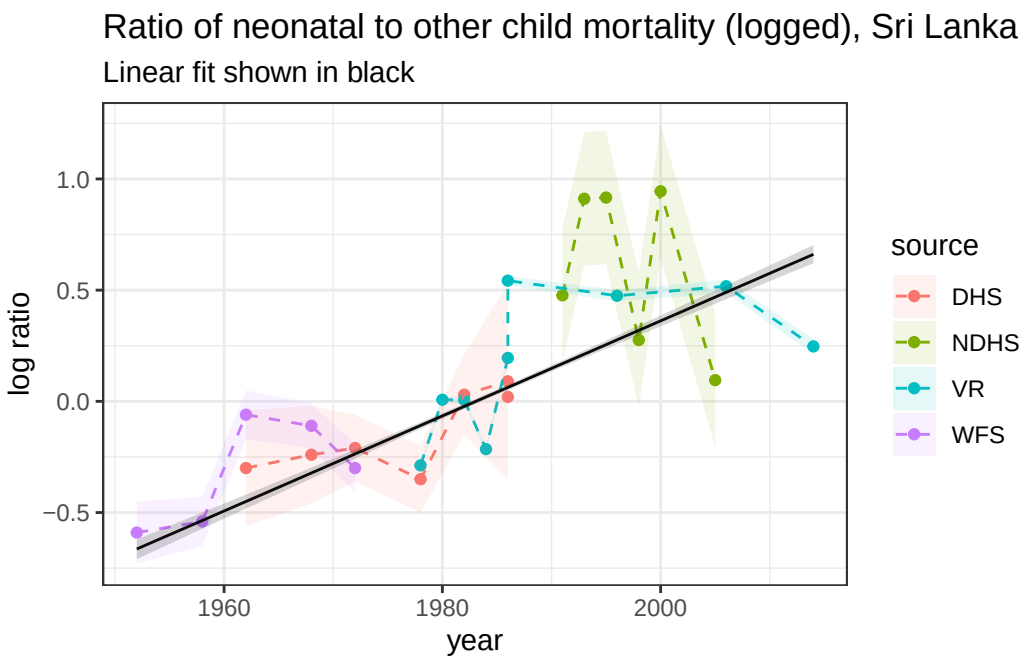
Extract the results:

```
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")
```



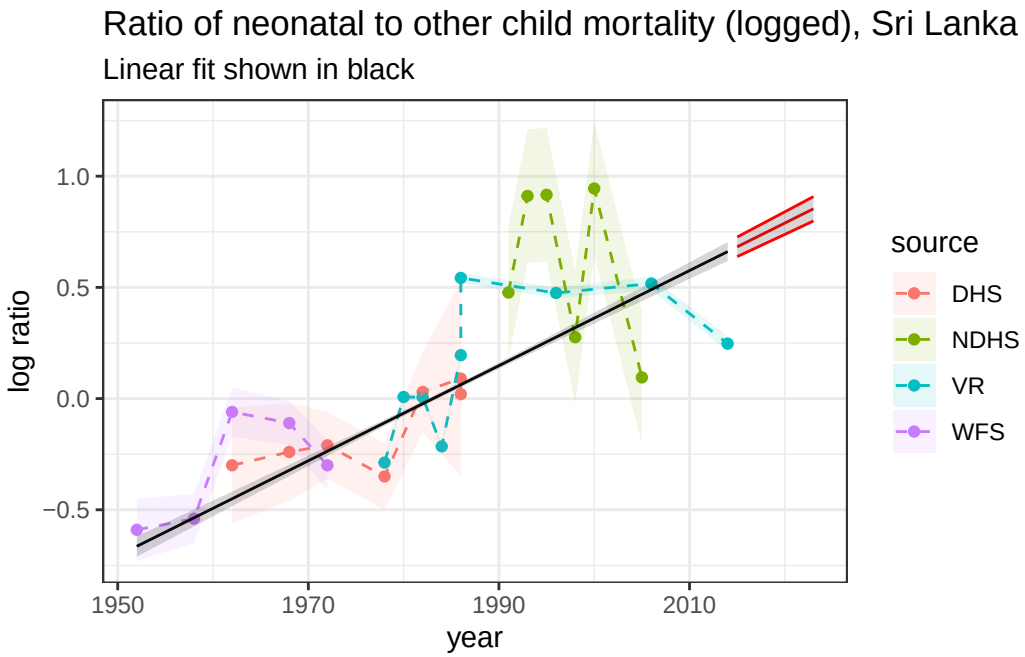
Question 1

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)

mod2 <- stan(data = stan_data,
             file = here("lka_linear1.stan"),
             refresh = 0,
            verbose = FALSE)
```

```
res <- mod2 |>
  gather_draws(mu[t]) |>
  median_qi() |>
  mutate(year = years[t], model = "Linear")

res_p <- mod2 |>
  gather_draws(mu_p[p]) |>
  median_qi() |>
  mutate(year = years[nyears]+p, model = "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 = 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, co
  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

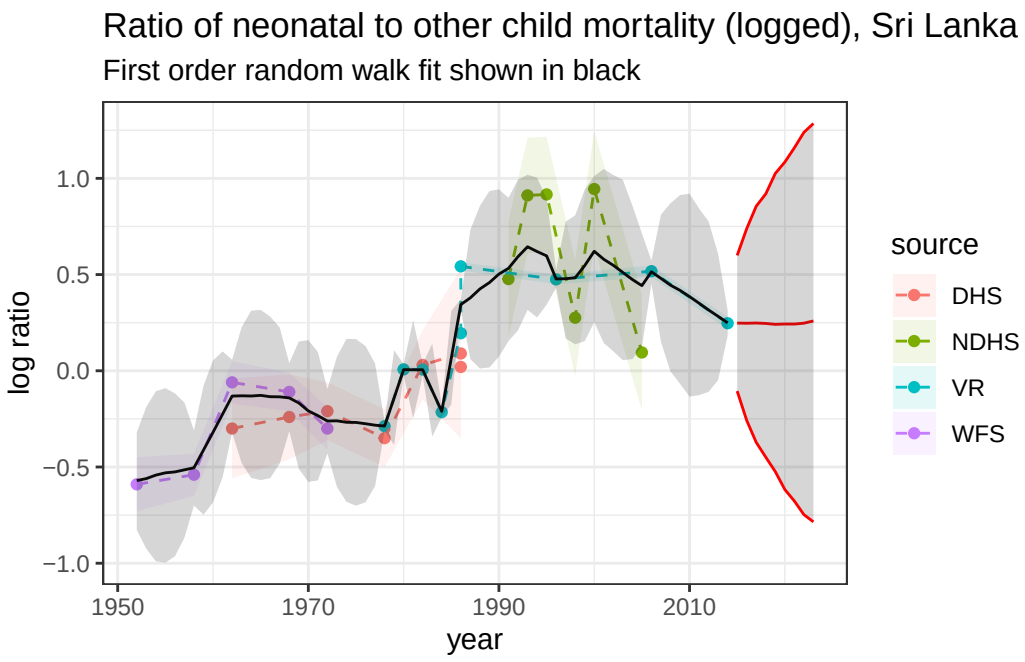
Code up and estimate a first order random walk model to fit to the Sri Lankan data, taking into account measurement error, and project out to 2023.

```
mod3 <- stan(data = stan_data,
             file = here("lka_linear2.stan"),
             refresh = 0,
            verbose = FALSE)
```

```
res2 <- mod3 |>
  gather_draws(mu[t]) |>
  median_qi() |>
  mutate(year = years[t], model = "1st Order")

res2_p <- mod3 |>
  gather_draws(mu_p[p]) |>
  median_qi() |>
```

```
  mutate(year = years[nyears]+p, model = "1st Order")
```

```
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, c
  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")
```



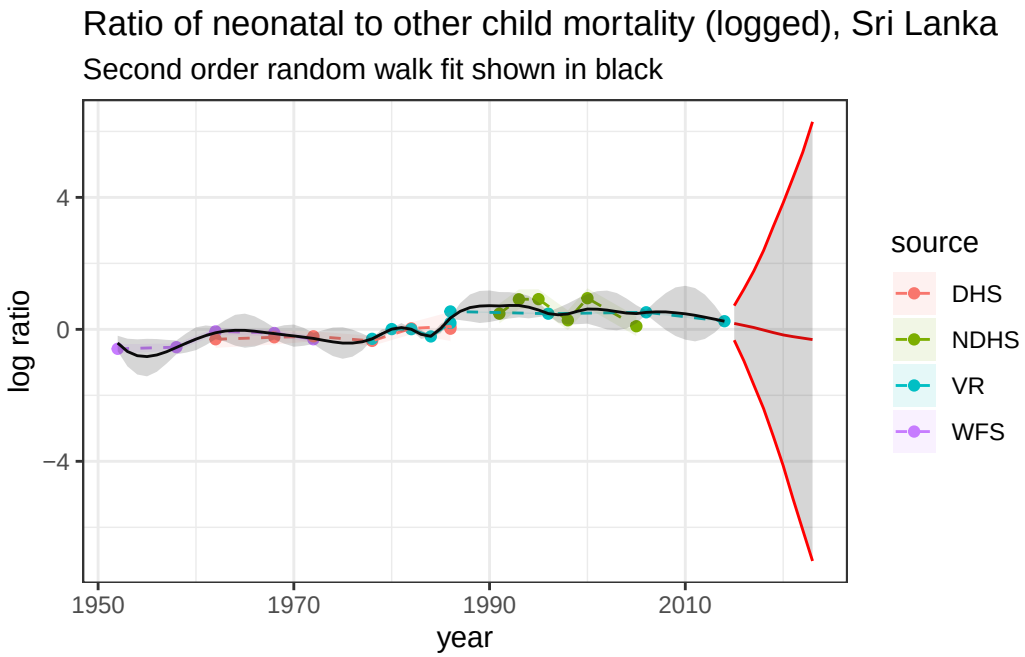
Question 3

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

```
mod4 <- stan(data = stan_data,
             file = here("lka_linear3.stan"),
             refresh = 0,
            verbose = FALSE)
```

```
res3 <- mod4 |>
  gather_draws(mu[t]) |>
  median_qi() |>
  mutate(year = years[t], model = "2nd Order")

res3_p <- mod4 |>
  gather_draws(mu_p[p]) |>
  median_qi() |>
  mutate(year = years[nyears]+p, model = "2nd Order")
```

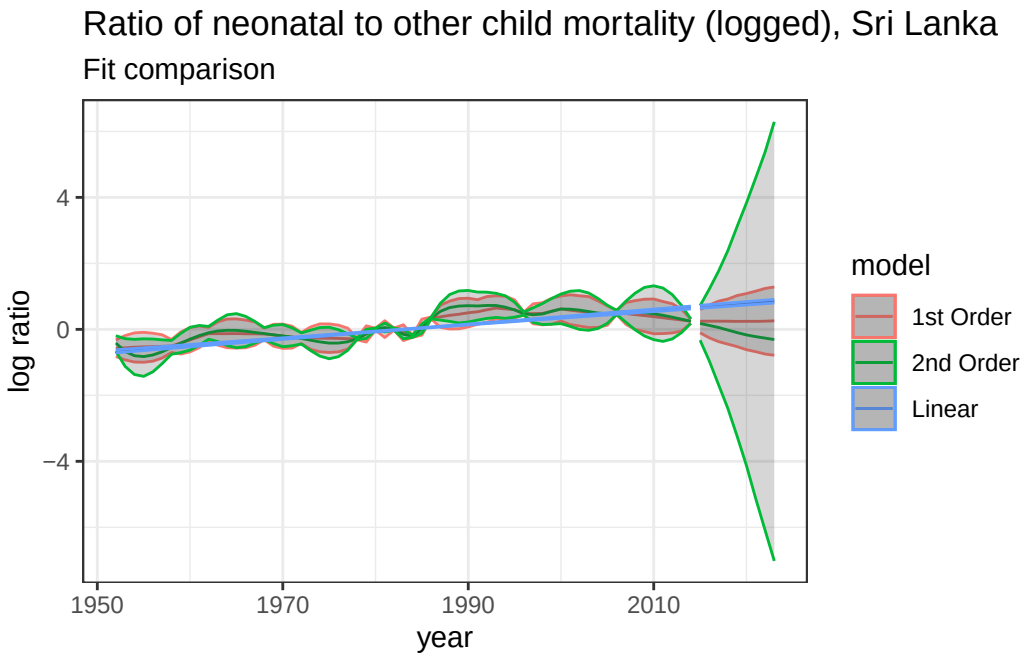
```
ggplot(lka, aes(year, logit_ratio)) +
  geom_point(aes( color = source)) +
  geom_line(aes( color = source), lty = 2) +
  geom_ribbon(aes(ymin = logit_ratio - se,
                  ymax = logit_ratio + se,
                  fill =  source), alpha = 0.1) +
  theme_bw()+
  geom_line(data = res3, aes(year, .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, c
  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.

```
res <- bind_rows(res, res2, res3)
res_p <- bind_rows(res_p, res2_p, res3_p)

ggplot(res, aes(year, .value, color = model)) +
  geom_line() +
  geom_ribbon(aes(y = .value, ymin = .lower, ymax = .upper, color = model), alpha = 0.2)+
  geom_line(data = res_p, aes(year, .value, color = model)) +
  geom_ribbon(data = res_p, aes(y = .value, ymin = .lower, ymax = .upper, color = model),
  theme_bw() +

  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "Fit comparison")
```



Question 5

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

VR data means data from that source

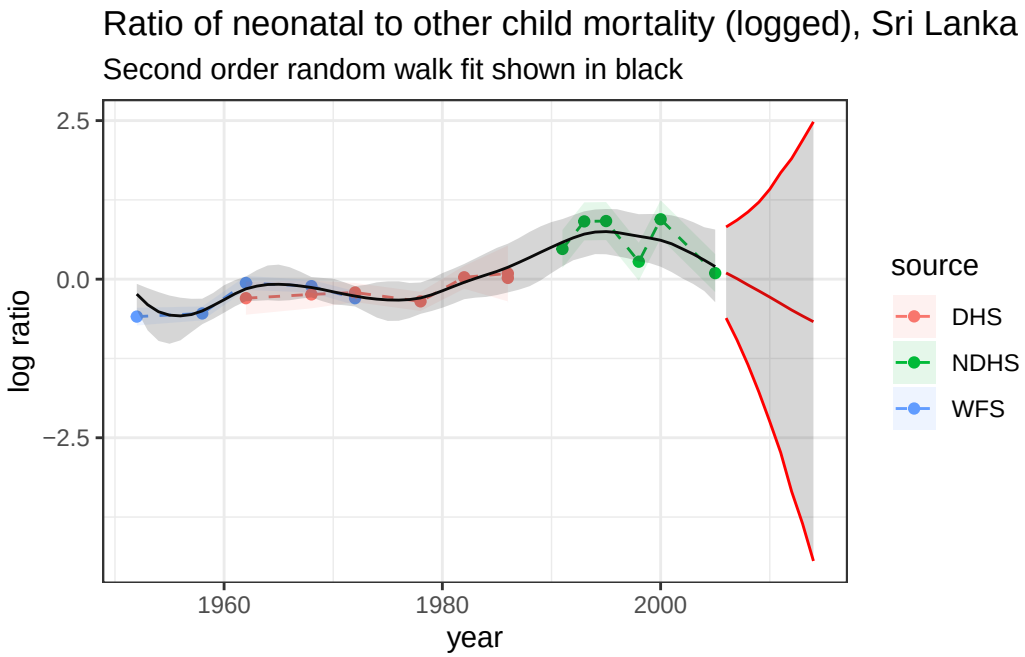
We can see that the uncertainty is much more stable in-data and reduced by close to half for projections. The exclusion of VR data means that there are some extra years with missing data. However, projection uncertainty is still decreased because VR data differs sgnificanlty from the other sources, and mainly NDHS.

```
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$logit_ratio, year_i = observed_years - years[1]+1,
                  T = nyears, years = years, N = length(observed_years), se = lka_no_vr$se
```

```
mod5 <- stan(data = stan_data,
             file = here("lka_linear3.stan"),
             refresh = 0,
            verbose = FALSE)
```

```
res4 <- mod5 |>
  gather_draws(mu[t]) |>
  median_qi() |>
  mutate(year = years[t])

res4_p <- mod5 |>
  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), alpha = 0.2, c
  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 6

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

I think the last model shown (question 5). It better captures the dynamics of the data and provides a reasonable, non-static, estimate for future projections. Removing VR data really helps contain uncertainty, maintaining projections within more reasonable boundaries.