Lab 10

Yaqi Shi, 1003813180

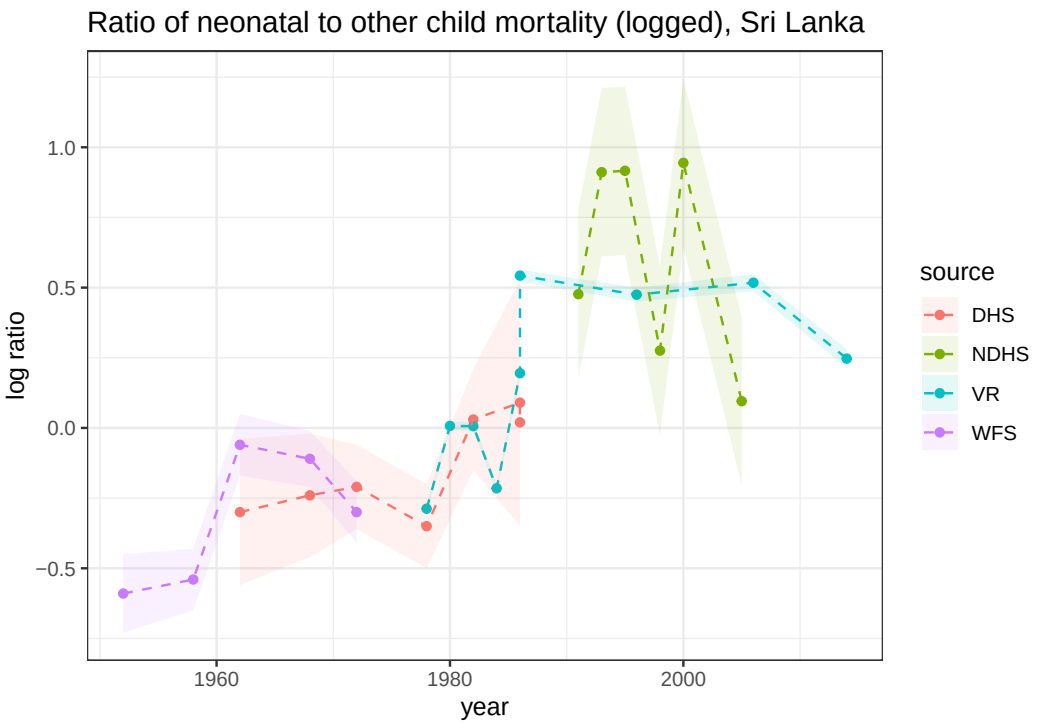
2023-03-26

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)

# Compose the data for stan
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)
```

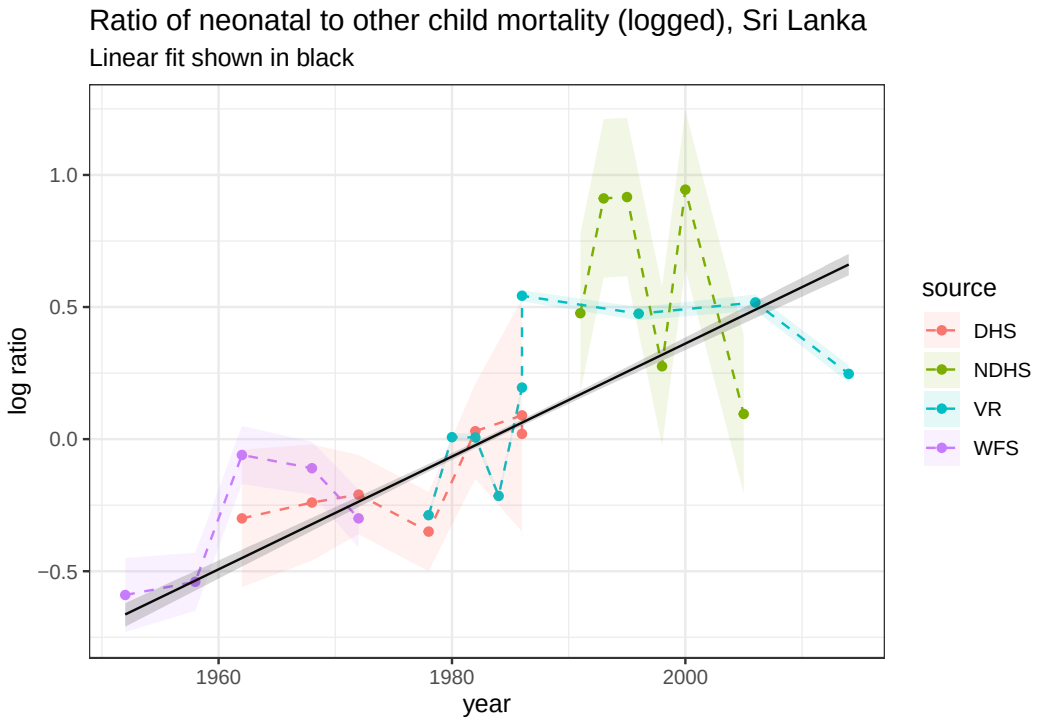
```
# Fit the model
mod <- stan(data = stan_data,
            file = "lka_linear_me.stan")
```

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) +
  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.

Answer

Here we run the model with generated quantities to project the rate till 2023.

```
# Compose the data for stan
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)
```

```
# Fit the model
mod2 <- stan(data = stan_data,
             file = "lab10_1.stan")
```

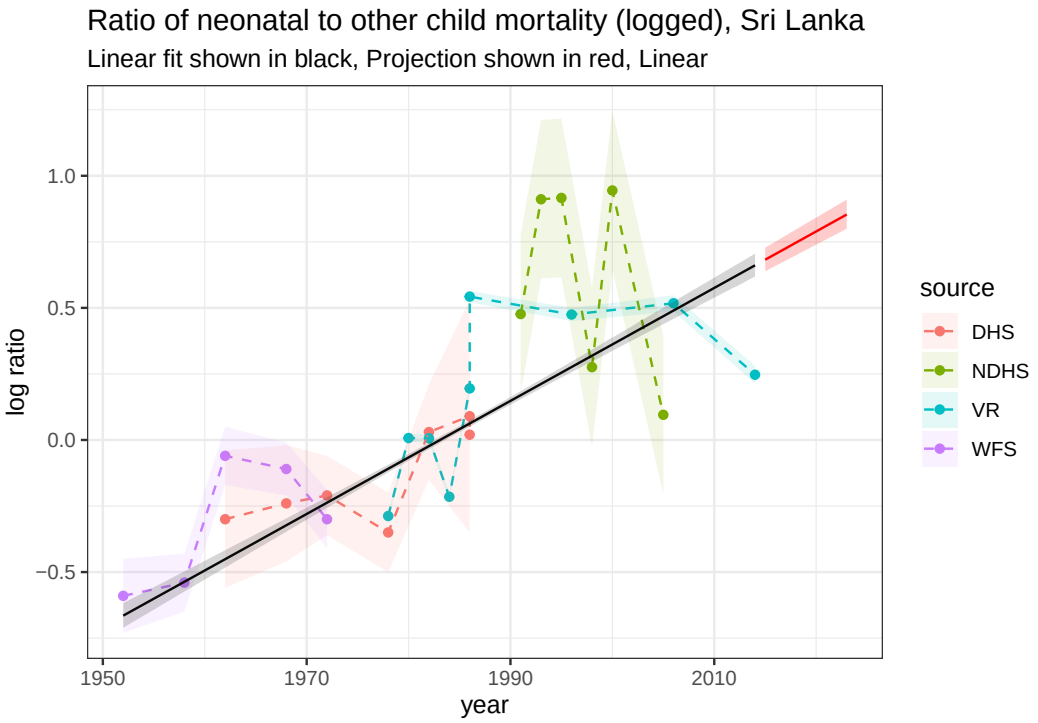
Extract the results:

```
res2 <- mod2 %>%
  gather_draws(mu[t])%>%
  median_qi() %>%
  mutate(year = years[t])

res_p2 <-  mod2 %>%
  gather_draws(mu_p[p])%>%
  median_qi() %>%
  mutate(year = years[nyears]+p)
```

Plot the resulting projections on a graph similar as before:

```
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 = res_p2, aes(year, .value), col = "red") +
  geom_ribbon(data = res_p2, 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 shown in red, Linear")
```



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.

Answer

Here is the model fit, details of the model is included in stan file:

```
mod3 <- stan(data = stan_data,
             file = "lab10_2.stan")
```

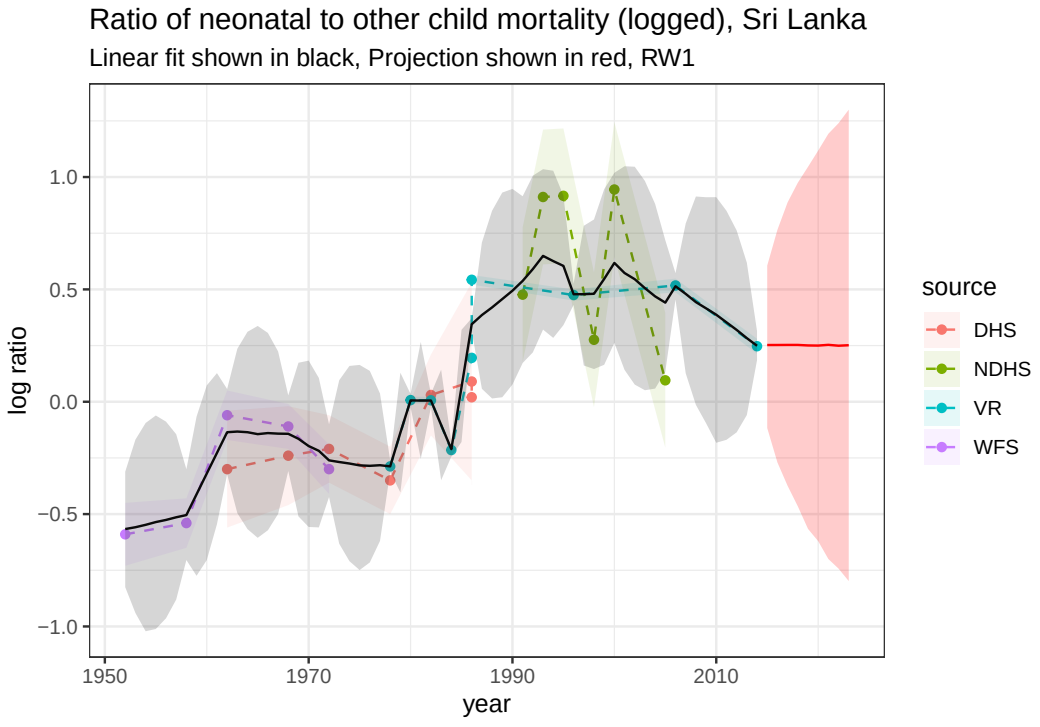
Extract the results:

```
res3 <- mod3 %>%
  gather_draws(mu[t])%>%
  median_qi() %>%
  mutate(year = years[t])

res_p3 <-  mod3 %>%
  gather_draws(mu_p[p])%>%
  median_qi() %>%
  mutate(year = years[nyears]+p)
```

Plot the model fit and projection to 2023 in a similar style:

```
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 = res_p3, aes(year, .value), col = "red") +
  geom_ribbon(data = res_p3, 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 shown in red, RW1")
```



Question 3

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

Answer

Here is the model fit, details of the model is included in stan file:

```
mod4 <- stan(data = stan_data,
             file = "lab10_3.stan")
```

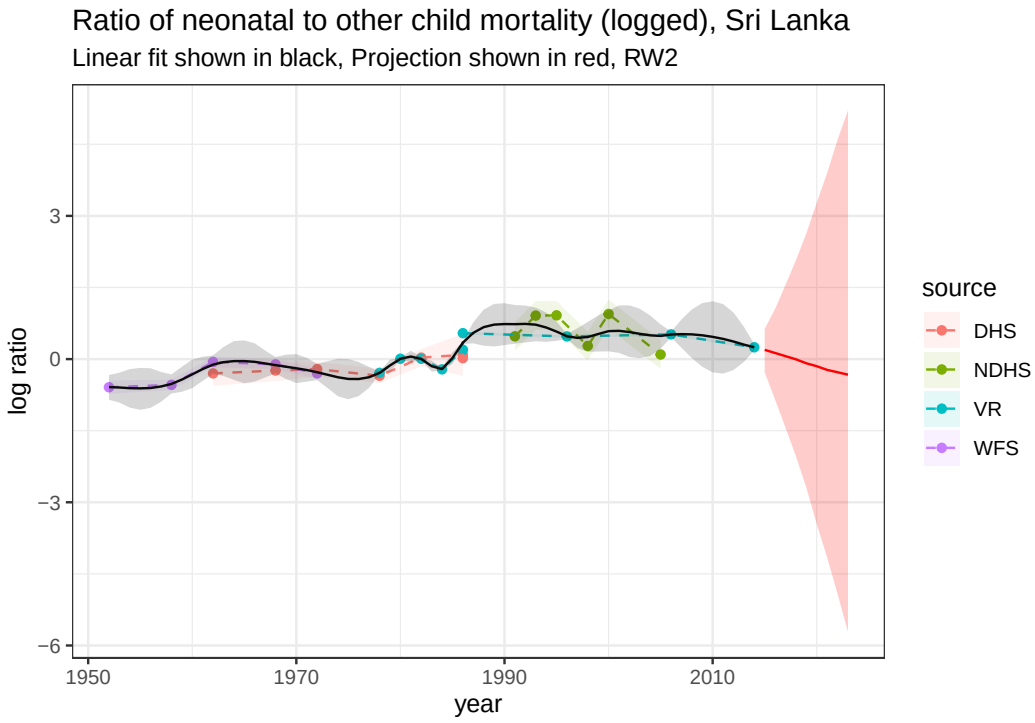
Extract the results:

```
res4 <- mod4 %>%
  gather_draws(mu[t])%>%
  median_qi() %>%
  mutate(year = years[t])

res_p4 <-  mod4 %>%
  gather_draws(mu_p[p])%>%
  median_qi() %>%
  mutate(year = years[nyears]+p)
```

Plot the model fit and projection to 2023 in a similar style:

```
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 = res4, aes(year, .value)) +
  geom_ribbon(data = res4, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2)+
  geom_line(data = res_p4, aes(year, .value), col = "red") +
  geom_ribbon(data = res_p4, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2, fill = "red")+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "Linear fit shown in black, Projection shown in red, RW2")
```



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.

Answer

I have run both model and now I plot the estimates including the projection to 2023 in the same plot below:

The red line and band corresponds to the linear fit.

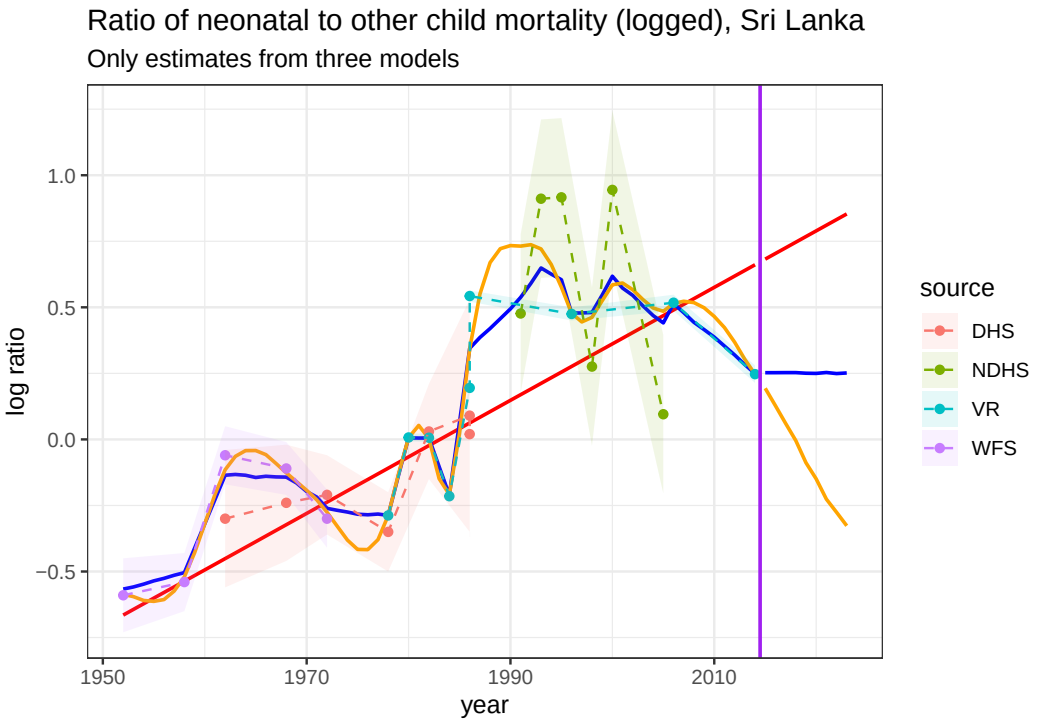
The blue line and band corresponds to the RW 1 model.

The orange line and band corresponds to the RW 2 model.

The purple line separates the fit and the projection.

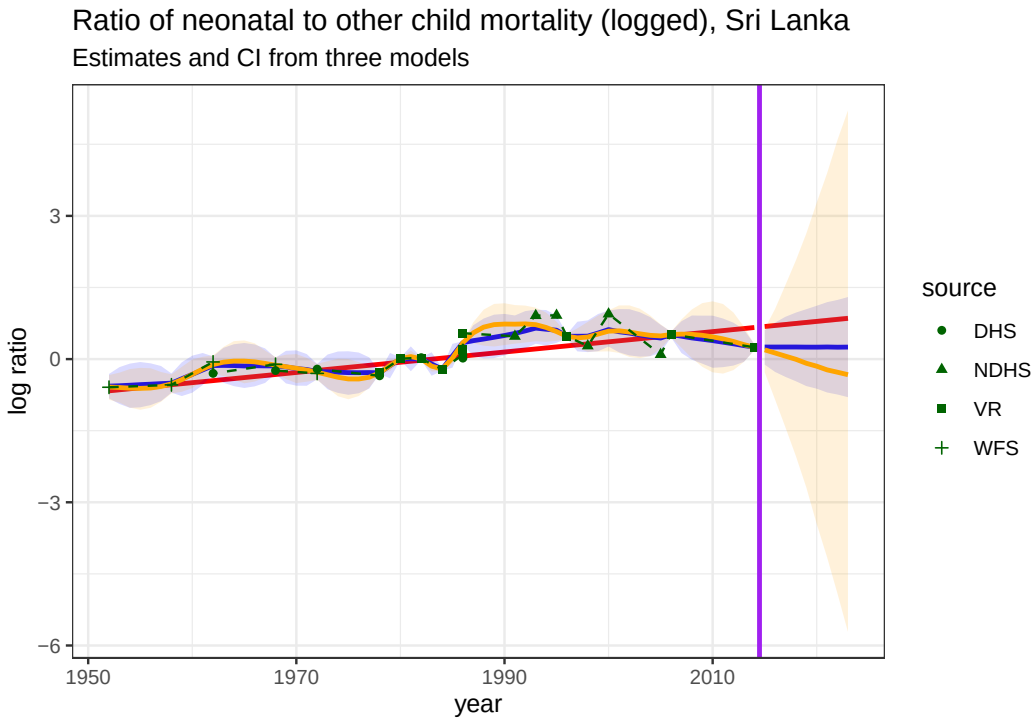
```
ggplot(lka, aes(year, logit_ratio)) +
  theme_bw()+
  geom_line(data = res2, aes(year, .value), col = "red", lwd = 0.75) +
  geom_line(data = res_p2, aes(year, .value), col = "red", lwd = 0.75) +
  geom_line(data = res3, aes(year, .value), col = "blue", lwd = 0.75) +
  geom_line(data = res_p3, aes(year, .value), col = "blue", lwd = 0.75) +
  geom_line(data = res4, aes(year, .value), col = "orange", lwd = 0.75) +
  geom_line(data = res_p4, aes(year, .value), col = "orange", lwd = 0.75) +
  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) +
```

```
  geom_vline(xintercept = 2014.5, color = "purple", lwd = 0.75)+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "Only estimates from three models")
```



In this plot, we include the credible interval into the plot and the green dots represents the data with different sources (different shape).

```
ggplot(lka, aes(year, logit_ratio)) +
  theme_bw()+
  geom_line(data = res2, aes(year, .value), col = "red", lwd = 1) +
  geom_ribbon(data = res2, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2, fill = "red")+
  geom_line(data = res_p2, aes(year, .value), col = "red", lwd = 1) +
  geom_ribbon(data = res_p2, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2, fill = "red")+
  geom_line(data = res3, aes(year, .value), col = "blue", lwd = 1) +
  geom_ribbon(data = res3, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.15, fill = "blue")+
  geom_line(data = res_p3, aes(year, .value), col = "blue", lwd = 1) +
  geom_ribbon(data = res_p3, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.15, fill = "blue"
  geom_line(data = res4, aes(year, .value), col = "orange", lwd = 1) +
  geom_ribbon(data = res4, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.15, fill = "orange"
  geom_line(data = res_p4, aes(year, .value), col = "orange", lwd = 1) +
  geom_ribbon(data = res_p4, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.15, fill = "orang
  geom_point(aes(shape = source), color = "darkgreen") +
  geom_line(lty = 2, color = "darkgreen") +
  geom_vline(xintercept = 2014.5, color = "purple", lwd = 1)+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "Estimates and CI from three models")
```



Comment:

Here we can see that the RW2 model have a larger CI in the prediction part compared to the linear fit and the RW1 fit. Both RW1 and RW2 shows are closer to the data compared to the linear fit. RW2 can fit the data in a more smooth way and it can catch the trends in the data. For the prediction section, RW1 model gives a flat prediction but RW2 model preserves the trend in the data.

Question 5

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

Answer

We first fit the RW2 model without VR data:

```
# Filter the data
lka <- lka %>%
  filter(source != "VR")

observed_years <- lka$year
years <- min(observed_years):max(observed_years)
nyears <- length(years)

# Compose the data for stan
stan_data1 <- 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 = 18)
```

```
mod5 <- stan(data = stan_data1,
             file = "lab10_3.stan")
```

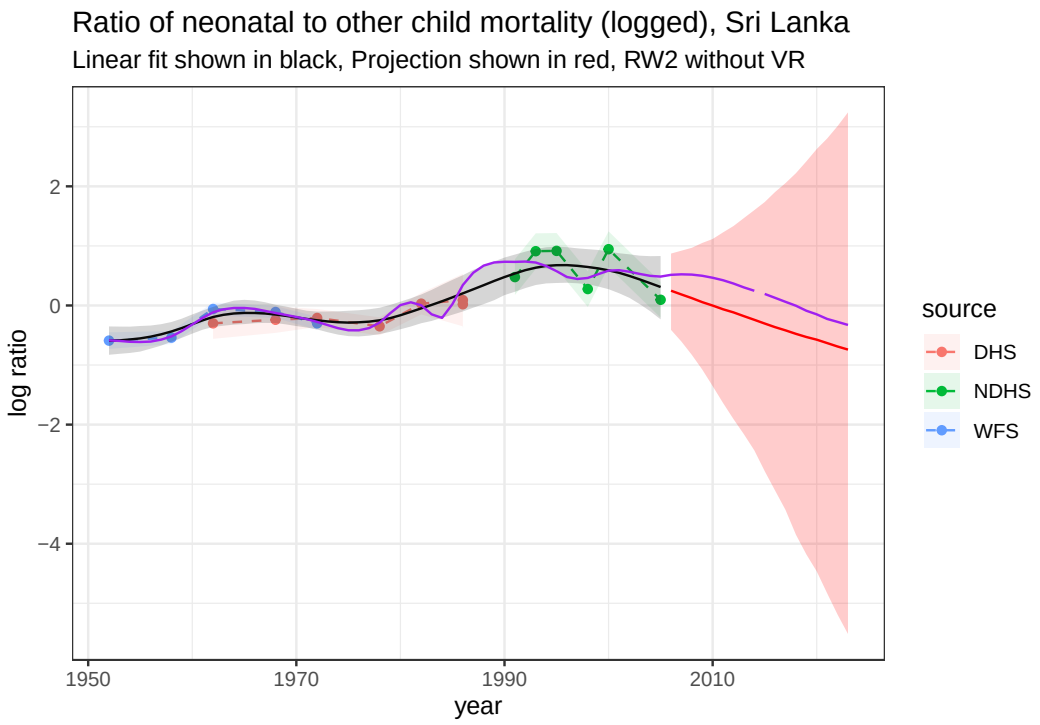
Extract the results:

```
res5 <- mod5 %>%
  gather_draws(mu[t])%>%
  median_qi() %>%
  mutate(year = years[t])

res_p5 <-  mod5 %>%
  gather_draws(mu_p[p])%>%
  median_qi() %>%
  mutate(year = years[nyears]+p)
```

Plot the model fit and projection to 2023 in a similar style:

```
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 = res5, aes(year, .value)) +
  geom_ribbon(data = res5, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2)+
  geom_line(data = res_p5, aes(year, .value), col = "red") +
  geom_ribbon(data = res_p5, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2, fill = "red")+
  geom_line(data = res4, aes(year, .value), col = "purple") +
  geom_line(data = res_p4, aes(year, .value), col = "purple") +
  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "Linear fit shown in black, Projection shown in red, RW2 without VR")
```



Comment:

The major difference is that without VR data, we only have data up until 2005 instead of 2014. The number of observations reduced from 27 to 18. Interestingly, the model is still able to catch the major trend of the rates. Also, since we have less data, the fitted line has less bumps in the middle (around 1990). The predicted number is slightly smaller than the RW2 model prediction with VR data. I have the fit and projection from RW2 model with VR in the purple line in the plot above to compare the results. Another aspect is that the CI is more smooth in the model without 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.

Answer

I would choose the model4 - second order random walk model with VR data (RW2 model with full data).

The reason is that: compared to the linear model, it fits the data better; compared to the RW1 model, it preserves the trend thus gives better projections. Thus it is preferred over the linear model and the RW1 model. Even though, without VR data, the fit and projections are more smooth and have less overlap data points (two rates for the same year), the resulting CI for the projection is smaller, however, there is no reason to exclude the information carried from VR data. VR data provides more information for years from 2005 to 2014 and it helps validate the trend in previous years where the data overlaps. I wouldn't exclude the data easily unless there are valid reasons to prove that the data is from a unreliable source or there are errors in the data. Thus I will choose model 4 (RW2 with full data) to be the most appropriate model.